MOTION AND EMOTION: SEMANTIC KNOWLEDGE FOR HOLLYWOOD FILM INDEXING

WANG HEE LIN

(*B.Eng.(Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRICAL AND

COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2007

ACKNOWLEDGEMENTS

This thesis would not have been able to take form without the help and assistance of many. I would like to express my heartfelt gratitude for the following persons and also to many others not mentioned here by name, for their invaluable advice, support and friendship in making this thesis possible.

Much gratitude is owned to my thesis supervisor, Assoc. Prof. Cheong Loong Fah, who provided both the theme of this thesis and the research opportunity for me. He has taught me much of the knowledge indispensable to research methodology. He has clarified my thought processes, built up my research experience and guided my direction more than anyone else. He has granted me much freedom in exploring the possibilities, without failing to provide valuable guidance along the way.

My reporting officer at I2R, Dr. Yau Wei Yun, must be thanked for his kind understanding and encouragement during the process of completing this thesis.

My heartfelt thanks to all the lab mates and FYP friends whom I have ever crossed path with during my stint at the Vision and Image Processing Lab, for their enriching friendship, assistance and exchange of ideas. In particular, I like to thank Wang Yong during my period of collaboration with him, as well as my fellow travelers Litt Teen, Shimiao, Chuanxin and Weijia, who brought me much joy with their companionship, and Francis for his assistance.

Finally, I cannot be more blessed by the presence of my mother, father, sister and grandpa for their unfailing love, encouragement and support in this endeavor in so many ways, for surely this thesis is dedicated to you for your wonderful love always.

TABLE OF CONTENTS

SUMMARY

In this thesis, we investigate and propose novel frameworks and methods for the computation of certain higher level semantic knowledge from Hollywood domain multimedia. More specifically, we focus on understanding and recovering the affective nature, as well as certain cinematographically significant semantics through the use of motion, from Hollywood movies.

Though the audience relates to Hollywood movies chiefly through the affective aspect, its imprecise nature has hitherto impeded more sophisticated automatic affective understanding of Hollywood multimedia. We have therefore set forth a principled framework based on both psychology and cinematography to understand and aid in classifying the affective content of Hollywood productions at the movie and scene level.

With the resultant framework, we derived a multitude of useful low-level audio and visual cues, which are combined to compute probabilistic and accurate affective descriptions of movie content. We show that the framework serves to extend our understanding for automatic affective classification. Unlike previous approaches, scenes from entire movies, as opposed to hand picked scenes or segments, are used for testing. Furthermore, the proposed emotional categories, instead of being chosen for ease of classification, are comprehensive and chosen on a logical basis.

Recognizing that motion plays an extremely important role in the process of directing and fleshing out a story on Hollywood movies, we investigate the relationship between motion and higher level cinema semantics, especially through the

philosophy of film directing grammar. To facilitate such studies, we have developed a motion segmentation algorithm robust enough to work well under the diverse circumstances encountered in Hollywood multimedia. In contrast to other related works, this algorithm is designed with the intrinsic ability to model simple foreground/background depth relationships, directly enhancing segmentation accuracy.

In comparison to the well behaved directing format for sports domain, shot semantics from Hollywood, at least at a sufficiently high and interesting level, are far more complex. Hence we have exploited constraints inherent in directing grammar to construct a well-thought-out and coherent directing semantics taxonomy, to aid in indexing directing semantics.

With the motion segmentation algorithm and semantics taxonomy, we have successfully recovered and indexed many types of semantic events. One example is the detection of both the panning establishment and panning tracking shot, which share the same motion characteristics but are actually semantically different. We demonstrate on Hollywood video corpus that motion alone can effectively recover much semantics useful for video management and processing applications.

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

1.1 Introduction

Motion pictures occupy a central position in popular entertainment. As a rich medium able to capture the human senses (sight and sound), staged dramatic renditions, movies, miniseries and dramas enjoy immense popularity in the modern age. The latest updated statistics of IMDB (Internet Movie Data Base) states that a mind-boggling 315 thousand movies have been released [109] to date. Bearing in mind that the vast majority of those films are of Western origin, one can expect a literal explosion of movie production as the film-making industries of other cultures mature and the technical cost of film-making continues to drop.

At the same time, the internet has steadily boomed over the past decade to be major vehicle of video data delivery and online commence. Several search engines like Google and Yahoo, along with video communities such as Youtube and Netflix, have arisen as a logical and necessary response to index, organize and search for video data on the sprawling World Wide Web, whose information would otherwise be nearly inaccessible to the masses. IPTV (Internet Protocol Tele-Vision), an anytime anyplace internet global channel delivery service, has also started to boom. In a similar vein, the confluence of these two major developments has led to the unprecedented demand for search engines specifically tailored to search and analyze motion pictures in a customizable manner for indexing, highlighting, summarization, data-mining,

automated-editing, recommendation and ultimately retrieval. With such a vast potential for automated commercial applications to fulfill the requirements of the general consumer, commercial vendors and niche markets, the possibilities of exploration in this field seems tremendous. Due to the immense popularity of Hollywood movies and the exponentially growing access and demand for it, the Hollywood multimedia domain stands out simultaneously as a most challenging and yet rewarding domain for machine understanding and processing.

Thus in this thesis, we investigate the indexing of movie resources with semantic concepts, or semantic indexing, using two salient aspects of the Hollywood movie domain: motion and emotion. The strongest commonalities underlying these aspects that recommend them for this work are: 1) their inspiration from cinematography and 2) the high level of movie semantics recoverable from them. The first part of the thesis develops the theoretical framework for affective (emotional) classification and analysis of movies, something that due to its complexity has hitherto received little attention. The framework, which is based on integrating the fields of cinematography and psychology, is then used to deal with key issues surrounding machine affective understanding of movies and designing effective cues for implementation. The second part of the thesis explores the rich repertoire of semantics that can be computed from shots using motion based features and characteristics. Once again, the theoretical basis of the taxonomy for the recoverable semantics is grounded in cinematography. Additionally an intricate algorithmic framework, which involves motion segmentation, is presented to enable the recovery of semantics, thus demonstrating the efficacy of motion for movie indexing.

The rest of the chapter starts with a brief explanation of semantic indexing. We then give a brief overview of prior works and also explain in more detail the two aspects of semantic indexing investigated: for 1) affective understanding of film and 2) shot semantics using motion respectively. Finally a summary of the contribution of the thesis is presented, followed by the thesis organization.

1.2 Semantic Indexing

To anchor the discussion and overview on semantic recovery for Hollywood movie indexing, some commonly used terms are defined here. In this work, a *document* is taken to be self-contained and coherent data that can be expressed in the digital format, with the more common forms being a story text file, image or song. Most types of documents are naturally organized around a hierarchical structure, where the more basic units of information are integrated together to form more complex units.

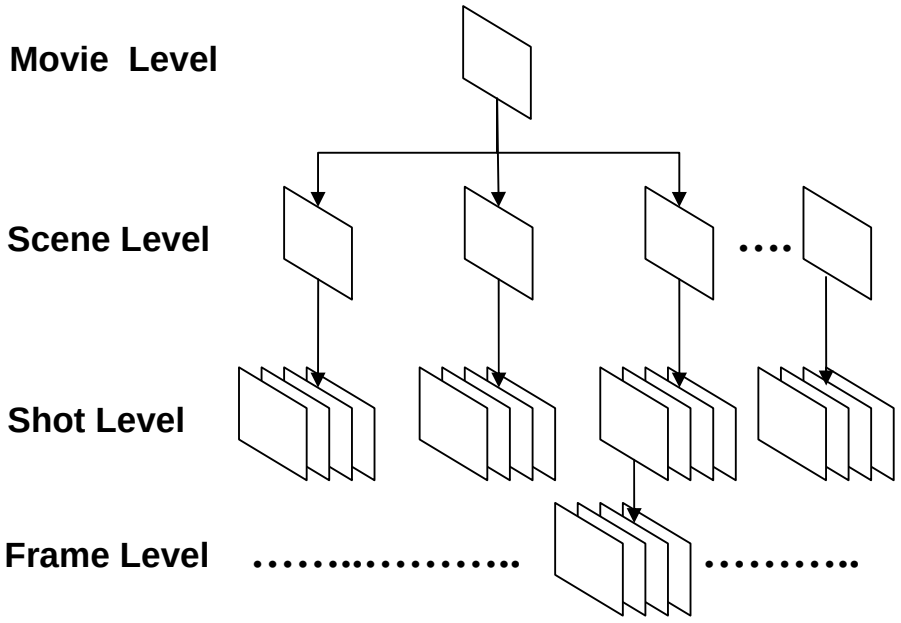


Figure 1.1 Hierarchy structure of a movie

Taking the analogy of a story, the individual words would be the basic units from which a more complex unit, such as the sentence, would be formed. Intuitively, sentences convey meanings or concepts which a user can relate to and are therefore interested in (high-level); we use the word *semantics* as a generic label for high level meanings and concepts. Individual words, on the other hand, cannot express ideas of sufficient interest in the absence of context (low-level). The movie possesses a similar hierarchy of information units, or levels of abstraction, which in descending order of complexity are the movie, scene, shot and finally individual frame (Figure 1.1). In reality, what data exactly constitutes as semantics is rather application and user dependent, and depends strongly upon the choice of level of abstraction.

Whatever the case may be, the process of extracting semantics can be simplified to an indexing process. At its most fundamental level, this process of document semantic indexing can be thus described: locating occurrences of similarity within the document based on similarity with pre-defined semantic models. This is the main reason why indexing is such a critical capability in the exploitation of movie resources. With this capability, vast movie resources can be automatically classified and organized according to personalized and innovative semantic labels that manual annotation cannot possibly anticipate or accommodate. This greatly enhances the browsing experience by paring down an unmanageably large list to a short list of well chosen candidates.

However semantic indexing of movies faces two tenacious problems. Firstly, as opposed to most present indexing works which deal with narrow domains, the movie domain has practically unlimited “variability in all relevant aspects of its appearance” Smeulders et al [108]. This implies the classification system must be carefully

designed to ensure that indexed semantic content remains well defined. Secondly, the greatest challenge to semantic indexing lies in bridging the *semantic gap*, which describes the apparent lack of relationship between low-level cues that are easier to compute and high-level semantics, which are more interesting. We note that indexing is more difficult than retrieval, which only needs to locate similarities with a given example within a document without any need for classification models.

1.3 Brief Overview of Semantic Recovery Works

Semantic recovery works are generally characterized according to two main aspects: by the level or the type of semantics being indexed. Because document level has a definite structure (Figure 1.2), the overview is organized according to the document level at which semantics are recovered.

The topmost level is the genre of a video document, which is the broad class a video document belongs to (e.g. sport, news, cartoon). Genre types that are relatively well defined and commonly recognized (especially program type) are popular amongst researchers. A brief history of genre classification shows that the genres tackled include: cartoon, news, commercial, music and sport [34][112]. At a slightly finer resolution, movies have been classified into genres (the rough category it belongs to) [11] and sports footage classified into the exact sport [111].

The next level is the scene level, which comprises of a consecutive series of shots. Semantics that hold coherent meaning at this level are plot elements, themes and location. Hence some works have attempted to detect the scene boundaries in order to recover the movie structure [2]. Recently the affective content of scenes has begun to receive attention from the indexing community. Pfeiffer used acoustic data mode to

detect for violent scenes in the MoCA (Movie Content Analysis) [106] project while Kang tried to recognize scene emotions using HMMs [10]. Note that our affective understanding work takes place at the scene level.

The next lower level of semantics belongs to the shot level, where certain short duration "behavior" events take place, thus there are still some conceivable applications where shot level analysis and retrieval is called for. Eickeler tried to differentiate between classes of news shots: anchor shot, interview and report [114]. Haering [83] carried out event detection and applied it to hunts in wildlife videos. In the sports domain, Lazarescu [75] analyzed football videos for sports events like different football plays while Duan detected goal scoring using video and motion features [77]. Our shot semantics from motion work takes places at the shot level.

As the building blocks of events, objects are conceptually the lowest level of semantics that users are probably interested. Due to its specificity, object indexing usually requires very strong *a priori* knowledge, encoded in the form of an object model. One of the most common objects to be detected or classified is the human face, by Kobla [110] and in the Name-It project at CMU [113].

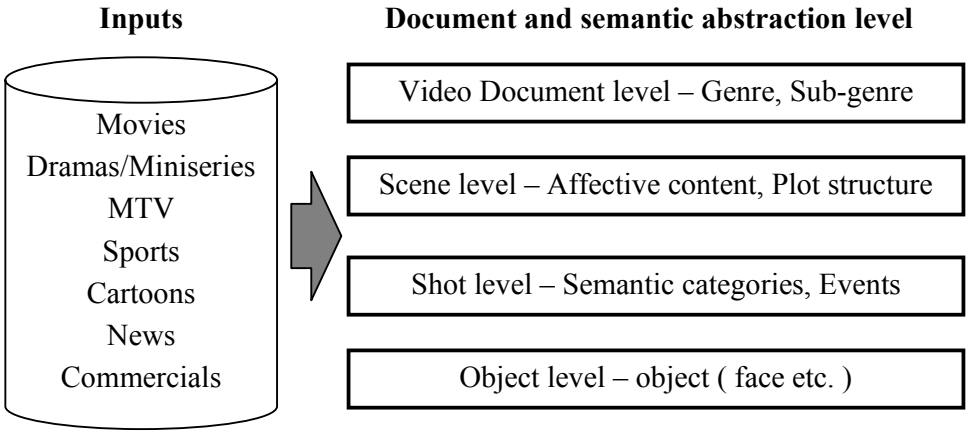


Figure 1.2 Semantic abstraction level for a movie.

1.4 Affective Understanding of Film

Indisputably, the affective component is a major and universal facet of the movie experience, and serves as an excellent candidate for indexing movie material. Besides the obvious benefit of indexing, automated affective understanding of film has the potential to lead to a new emotion-based approach towards other hotly researched topics, including video summarization, highlighting and querying. This paves the way for even more exciting but unexplored applications, such as movie ranking and personalized automated movie recommendation.

Film does not develop or exist independently of human psychology and culture, and the underlying principles behind many aspects of film grammar become clearer from a psychological perspective. In our work, we recognize and establish the intimate relationship between cinematography and psychology for affective understanding of film, as well as the benefits that an integration of the insights from both these fields will bring. Consequently we have used the methods and theories from both fields on a complementary basis to develop the required conceptual framework and design the low-levels cues necessary for affective classification of Hollywood movies.

Because the movie structure naturally demarcates affective content at the scene level, we have chosen the scene as the basic unit for semantic affective extraction. We show that this information can actually be used to accurately classify the affective characteristics of movies at the higher document level. More significantly, we demonstrate the ability to infer the degree of different affective components in film, a step up in the sophistication level and usefulness, compared to classification alone.

1.5 Film Shot Semantics from Motion and Directing Grammar

Content based Visual Query (CBVQ) semantic indexing systems have recently come to appreciate that motion holds a reservoir of indexing information. This is most true of narrative videos like movies, where camera movement and object behavior are purposive and meaningfully directed to elucidate the intentions of the producer and aid the story flow. Guiding the director is a set of production rules on the relationships between shot semantics and motion, which are embodied in a body of informal knowledge known as film directing grammar.

In this work, we explicate the intimate multifaceted relationship that exists between film shot semantics and motion by appealing to directing grammar. Based on our insight that manipulation of viewer attention is what ultimately defines the directing semantics of a shot, we have formulated a novel edge-based MRF motion segmentation technique, with integrated occlusion handling, to capture the salient information of the attention manipulation process.

Directing grammar has also provided us with the framework to propose a coherent semantics taxonomy for film shots, and to design effective motion-based descriptors capable of mapping to high level semantics. Using both the motion segmentation algorithm and semantics taxonomy, we can recover semantics like director intent and possibly story structure from motion, which in turn directly aids film analysis, indexing, browsing and retrieval.

For this work, the shot, which is the only unit to comprise an uninterrupted flow of motion, is naturally adopted as the basic unit for study.

1.6 Summary of Contributions

Here we summarize the contributions of the thesis in point form:

Affective Understanding of Film

- Using psychology and cinematography to create a theoretical basis and framework for affective understanding of multimedia; exploring affective related issues.
- Deriving a set of useful audio-visual low level cues for affective classification of movie scenes, especially a probabilistic method of accurately extracting affective scene information from noisy movie audio.
- Investigate into the affective nature of movies and movie scenes.
- Demonstrating innovative affective based applications with good results.

Film Semantics from Motion

- Investigating and developing the use of cinematography as the theoretical basis for using motion exclusively to recover semantic level information from movie shots.
- Proposing a robust motion segmentation method capable of segmenting out video semantic objects (foreground and background) for use with Hollywood movies.
- Proposing an organization principle based on film directing elements and grounded in directing grammar to construct a well-formed and coherent film directing semantics taxonomy.
- Designing effective and robust descriptors to recover shot semantics using motion.
- Demonstrating the proposed framework with good classification results.

1.7 Thesis Organization

The rest of the thesis is organized as follows.

In Chapter 2, we investigate the affective aspect of Hollywood movies. We introduce the foundational methodologies, consisting of both of psychology and cinematography, on which the theoretical framework necessary for developing the rest of the affective work is based. Several inevitable issues arising from affective classification in Hollywood multimedia are discussed, particularly choosing an appropriate set of output emotional categories.

Chapter 3 builds upon the framework in the previous chapter to propose and justify a set of powerful audiovisual cues. A probabilistic inference mechanism based on the SVM is introduced, which produces the final probabilistic affective outputs. A comprehensive set of experiments are carried out, followed by a discussion of the results. Two applications of the affective classification framework are demonstrated.

Chapter 4 proposes a new motion segmentation algorithm specifically suited to the purpose of film semantics recovery from motion. Ways to overcome likely problems are discussed. The implementation, performance characteristics are explored in detail and experimental results are demonstrated.

Chapter 5 introduces a semantic taxonomy to classify shot semantics in the film domain using motion. We justify this taxonomy based on cinematography and in turn use it to formulate both the low level motion descriptors and the output semantic classes. Finally the experimental results for film indexing using the resultant framework are shown.

Chapter 6 concludes the thesis with its implications and potential future works.

CHAPTER II

AFFECTIVE UNDERSTANDING IN FILM

2.1 Introduction

With the increasingly vast repository of online movies and its attendant demand, there exists a compelling case to empower viewers with the ability to automatically analyze, index and organize these repositories, preferably according to highly personalized requirements and criteria. An eminently suitable criterion for such indexing and organization would be the affective or emotional aspect of movies, given its relevance and everyday familiarity. Endowing an automated system with such an affective understanding capability can lead to exciting applications that enhance existing classification systems such as movie genre. For instance, finer categories such as comedic and violent action movies can be distinguished, which would otherwise have been grouped together in the action category under the present genre classification.

With the ability to estimate the intensity of different emotions in a movie, a host of intriguing possibilities emerges, such as being able to rank just how “sad” or “frightening” a movie scene is. Taken to its logical end, this can lead to personalized affective machine reviewer applications, doing away with the limitations of predefined movie genres. In short, computable affective understanding promises a new emotion-

based approach towards currently investigated topics such as automated content summarization, recommendation and highlighting.

Surprisingly, immediately related works in affective classification of general domain multimedia have been few. While many works exist in the wider area of multimedia understanding, ranging from scene segmentation [2], sport structure analysis [3], event detection [4], semantic indexing in documentaries [5], sports highlight extraction [6], audio emotion indexing [7] to program type classification [34], literature in affective classification is sparse and recent. This state of affairs is mainly due to the seemingly inscrutable nature of emotions and the difficulty of bridging the affective gap [8], especially in this case where high level emotional labels are to be computed from low level cues.

Of works that deal with affectively-related issues, [9] computed the motion, shot cut density and pitch characteristics along the temporal dimension of movie clips from which emotion profiles known as “affect curves” are obtained in a 2D emotional space known as the Valence-Arousal space. [10] used visual characteristics and camera motion with Hidden Markov Models (HMM) separately at both the shot and scene level in an attempt to classify scenes depicting fear, happiness or sadness, while [11] proposed a mean-shift based clustering framework to classify film previews into genres such as action, comedy, horror or drama, according to a set of visual cues grounded in cinematography. [12] proposed Finite State Machines (FSM) with face detection and an audiovisual based activity index to model and distinguish between conversation, suspense and action scenes.

While these works have advanced research in affective classification, their output emotion categories in the affective context are somewhat ad hoc and incomplete

[10]-[12] ([9] does not use output emotions). Furthermore, the inputs treated by these works are previews [11] or handpicked scenes [10], which due to the prior manual filtering process, are biased by the aims and methods of the selectors. It remains to show whether these works can be readily extended to treat more emotions as well as to analyze complete movies. Crucially, the following important questions are left unaddressed: How should output emotion categories be chosen? And what should they actually be?

Thus in establishing a successful movie affective understanding system, we put forth, as our first contribution, a complementary approach grounded in the related fields of cinematography and psychology. This approach identifies a set of suitable output emotion categories which are chosen with clear reason, a more complicated task than it seems. The increase in the number and subtlety of these categories results in a more difficult, but also more comprehensive and meaningful classification. In contrast, besides having less complete output emotion categories, previous works are explicitly based on just one of the two fields. In the film affective context, they are thus constricted by the limited information and paradigms at their disposal. [9] employed only psychology and [11] cinematography, while [10] mentioned psychology briefly but proceeded solely based on the cinematographic basis.

For our second contribution, we develop from cinematographic and psychological considerations a set of effective audio-visual cues in the film affective context. Though low-level, some of these features can yield high-level information which helps to bridge the affective gap. For instance, we formulate a visual excitement feature that takes viewer feedback directly into account. Other useful features, which have not been employed in this context, are color energy, chroma difference, music

mode and the proportions of Music, Speech and Environ (MSE) audio. In particular, we propose a probabilistic based approach to extract movie audio affective information from each of the MSE channels in a more suitable and comprehensive manner than other film affective works. This is accomplished by splitting the audio analysis units according to cinematographic knowledge and processing each MSE channel differently to overcome confusing multiple-speaker presence and MSE mixing, amongst other challenges.

Due to the dominance of the "classical Hollywood cinema" in film [1], the scope of this work deals with automatically analyzing and classifying the affective content of Hollywood movie scenes, and in turn the entire movies. The scene, also known as the story or thematic unit, is chosen as the basic unit of analysis, because it conveys semantically coherent content, and is the primary unit of distinct phases of plot progression in film [1]. The notion of mise-en-scene, where the design of props and settings revolve around the scene, further enhances its potency [1]. Not surprisingly, it is usually the individual scenes that are most sharply etched in the collective memories of the cinema.

This chapter introduces the background and explores the fundamental issues of the work. Then we lay down the proposed complementary approach and demonstrate how it guides us in choosing the output emotional categories. Chapter 3 discusses the probabilistic framework, affective features designed for affective classification and presents the resultant system with experimental results. Figure 2.1 illustrates the scope covered by the affective work.

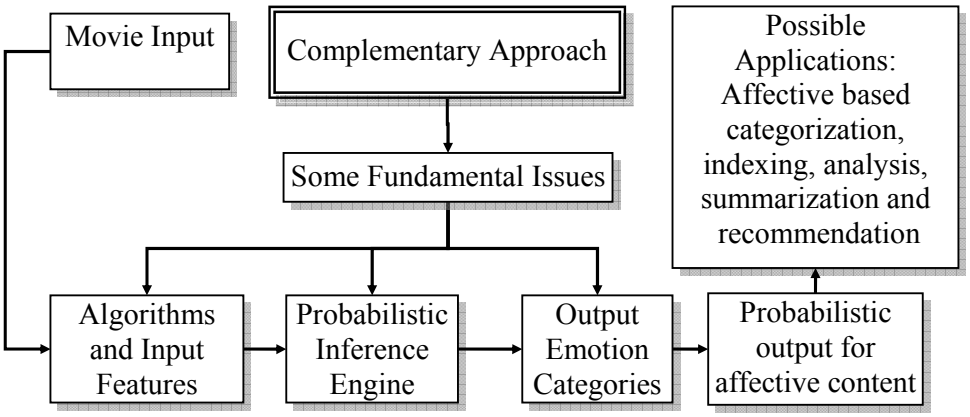


Figure 2.1 Illustration of scope covered in current work.

2.2 Review of Related Works

With such a small number of directly relevant works, we will review their frameworks, algorithms and experimental results in greater detail in the following paragraphs. Hanjalic [9] proposed a purely dimensional approach to affect by utilizing the emotion theory of Russell and Mehrabian [18], who investigated the nature of emotions and suggested that all emotions could be characterized completely by three basic primitive affective qualities. These qualities are respectively Valence (measure of pleasure), Arousal (measure of mental intensity or agitation) and finally Dominance (measure of psychological control). In order to characterize video content in the VAD space, certain computable audio-visual cues that can supposedly directly measure these qualities have been designed. However due to the limited utility of Dominance, only Valence and Arousal have been eventually adopted. The Arousal measure proposed is a weighted linear combination of the shot density, energy in higher frequency sound and the magnitude of motion vectors while the Valence measure comprised the audio pitch, which is valid only in places where speech is voiced. By computing Valence and

Arousal quantities along the temporal dimension of a movie clip, emotion profiles known as affect curves can be obtained in Valence-Arousal space. Hanjalic further suggested that the general location of such profiles could provide information about the dominant mood in clip. However he did not propose any output emotional classes which the affect curves could map to, nor are the two test video sequences used sufficient for drawing conclusions about the effectiveness of the algorithmic approach. Though attractive in its generality and simplicity, the overlap of fundamental emotions in such a VA space invalidates the exclusive use of VA space for affective classification, as will be shown later.

Kang [10] used camera motion and visual features to characterize every shot into either fear, happy, sad or normal emotions. The camera motion features consist of the motion type (pan, tilt etc.) and its magnitude while the visual features contain the amount of brightness as well as the proportion of “culture colors”, where “culture colors” (red, yellow, green, blue, purple, pink, orange, gray, black ,white etc.) are colors claimed to be imbued with cultural or emotional significance. Each shot is therefore described by a feature vector, which is compressed into a symbol via vector quantization. Hidden Markov Models (HMM) are then trained separately at both the shot and scene level in an attempt to classify both shots and scenes into the four emotions.

However such use of HMMs, especially for modeling emotions at the shot level, is fraught with problems for two reasons. Firstly, affect is not a well-defined concept at the shot level. Secondly, due to the first objection, the transitional probabilities for the HMMs are in turn not well-defined. Besides that, the testing data is inadequate (six 30 minutes worth of scenes, which usually last more than one minute

each). Furthermore, since the scenes are meticulously hand-picked, the introduction of bias is a likely probability that cannot be dismissed. The audio aspect has been neglected and finally, the output categories themselves are incomplete, and selected for the ease of classification.

Rasheed et al [11] considered the slightly different yet related problem of genre classification of movie previews into action, comedy, horror or drama. A set of exclusively visual cues grounded in cinematography are proposed to characterize every preview: namely average shot length, motion content, color variance and lighting key. The significant contribution of this work lay not so much in the cues itself as in the cinematographic foundations used to justify the cues, which provides a theoretical foundation for the cues. A mean-shift based clustering framework is finally used to cluster test previews into different genre membership clusters. However due to the fact that film previews are manual summaries of films used for the purpose of advertisement, only shots that epitomize the genre of the movie would be included, thus simplifying the film genre classification task tremendously, as opposed to affective classification for every single scene in a movie. The aural aspect of film, which plays a pivotal role in the affective experience, has also not been addressed.

Zhai et al [12] proposed using Finite State Machines (FSM) for classifying three different scene semantics (suspense, action and conversation). To accomplish this, two cues are extracted at the shot level. The first cue computes activity intensity, which is a weighted measure of the dominant motion vector, its variance and the mean audio intensity while the second cue detects the presence of human faces in every shot. Finite State Machines, which are a very specialized form of the Markov Model, are designed for each of these three semantics.

For instance, the FSM for detecting conversation specifies that there must be neighboring shots showing different faces and there must not be high activity shots. Each of these FSMs loops inevitably end with either an accept or reject node after certain sequences of shot types are encountered, regardless of the characteristics or number of shots remaining in a scene, which is very unrealistic. Furthermore, the endless variety with which such scenes can evolve is far beyond what simple hand-crafted FSMs can possibly capture. Finally the video corpus, at sixty clips and only involving these three types of scenes, is too small and unrepresentative.

The last work by Moncrieff [4] uniquely focused on examining localized sound energy patterns, or events, associated with high level affect experienced with horror films. Defining four types of sound events (composed of varying sound energy profiles) usually associated with the horror genre, the central idea of the work centers on inferring affects brought about by these well established sound energy patterns employed in audio tracks of horror films. Using window matching, locations corresponding to these events are detected. In order to ascertain the accuracy and effectiveness of these events in inferring the presence of "horror" scenes or film, statistics compiled from the six movies were analyzed. The results showed that sound event detection can distinguish between horror and non-horror films, as well as detecting horror scenes in horror films. This demonstrates the indicative power of sound at both the film and scene level, albeit only in the limited genre of horror.

As a broad comparison of our work with others, several main advantages emerge. Foremost among these, we have a set of output emotion categories that are theoretically better founded and thus more suited for affective film classification than the more ad hoc and limited emotional categories proposed by others. For instance, the

hand-crafted FSM used in [12] is inadequate for approximating the structural variety of many types of scenes, and [9] has not proposed any output categories. Secondly, we have exploited affective information for audio far more extensively than others, who concentrated on visual cues [9]-[12]. Thirdly, we have adopted a SVM based probabilistic inference engine capable of expressing beliefs in the affective components probabilistically instead of discretely [11], thus increasing output accuracy. Furthermore, this engine can be extended easily to accommodate more new affective features easily. Finally, we have not pre-selected our experimental data; and its size, at about two thousand scenes, is also larger than the next largest video corpus used [10] by about an order of magnitude.

2.3 Definition of a Scene

As the fundamental story unit around which the film is organized, the scene is invested with coherent and intelligible plot, causing the basic units of affect to be naturally demarcated along scene boundaries. Thus to facilitate affective scene classification, it is imperative to formulate a working definition for the scene that is consistent, appropriate and as objective as possible for the work. The term “scene” originates from a French classical theater term mise-en-scène, which literally means "put in the scene", and has a precise beginning and ending corresponding to the arrival and departure of characters [1]. Probably due to the inherent limitations of the theater and generally linear nature of the theater then, discerning scene boundaries was simpler. However in cinematography, the heavy use of editing as a technique (i.e. cutting) to form a narrative allowing events occurring in spatially different places to be portrayed as temporally parallel events - hence enabling the narrative to be

experienced in a more intuitive and non-linear manner - has blurred the meaning of scene boundaries as increasingly short duration cuts of different settings are interwoven one with another.

In [2], which seeks a computational approach to detect scene boundaries, the authors state that it is more appropriate to define scenes from the film maker's viewpoint and study cinematic devices to design an algorithmic solution. They used these set of guidelines to set up the ground truth and define scenes in their work:

1. When there are no multiple interwoven parallel actions, a change in location or time or both defines a scene change.

2. An establishing shot, though different in location to its corresponding scene, is considered part of that scene, as they are unified by dramatic incidence.

3. When parallel actions are present and interleaved, and there is a switch between one action to another, a scene boundary is marked if and only if the duration of that action is shown for at least 30 seconds. Their reasoning is that when an action is briefly shown, it serves more as a reminder than representation of any significant event. This implies that while supporting action shots may never make a scene, a long dominant action scene may possibly be broken into smaller scene units. An example is raised in the training scene of *The Matrix*, where a few short shots are inserted to show group members watching through a computer (i.e. a different locale), which should not be considered as making up a new scene.

4. Finally, a montage sequence which is formed by dynamic cutting, a technique where shots containing different spatial properties are rapidly joined together to convey a single dramatic event, constitutes a single scene. For example, in order to convey the desperate attempts of Carolyn Burnham to sell her house in *The*

American Beauty, the film maker joins many different shots of her showing different customers different parts of the house.

From experience, we find the above set of guidelines to be very useful. However since their primary motivation stems from finding a computable solution to scene segmentation, it is inevitable that the guidelines will not coincide with the definition of a scene that is more appropriate for affective scene classification. Although we fully concur with Guidelines 2 and 4, Guideline 1 throws up a question: what degree of change constitutes a change in time and location? For pursuing chase scenes, gradual slight locations/settings changes are natural. In a beginning scene from *Terminator*, Reese ran from the streets into a departmental store. There is at least a superficial change in setting, thus suggesting a scene boundary. Regardless, spectators will intuitively consider the street to store chase as one scene.

This judgment, we believe, is due to a very strong continuity and constancy of the characters, mood and semantics of the shots. Furthermore, the change in settings happened on a spatial continuity, with Reese running from one to the other location. Conversely, if there is a complete lack of continuity in the mood/semantics/characters between two groups of shots except for the setting, it will be very difficult for spectators to experience these shots as one scene. We believe that although the first rule generally holds, yet the judgment of sameness in time and setting depends much on the degree of semantic, character and mood (affect) continuity. Presently, this observation is incomputable and inevitably subjective to a certain extent. However we believe it does help to clarify the principles behind what really constitutes a scene.

Guideline 3 is strictly not true and has been violated many times, especially action films. In the famous last battle of *Star Wars: Episode I*, parallel narratives from

three vastly different places are tightly interwoven in a technique known as cross-cutting. These temporally parallel plotlines showing how three groups of people fighting a common enemy in as many places are largely independent of each other and do not serve as reminders in any sense of the word. It is also noted that some of their brief appearances in the narrative last longer than 30 seconds. However from the director's viewpoint, using such tight editing to produce this sense of mood/plot coherence and parallelism amongst all the shots can mean only one thing; these shots are meant to be experienced and remembered by the spectator as one long scene. We feel that the duration of 30 seconds to denote scene change is a good gauge, but can be set aside as long as the general pace and pattern of cross cutting is kept up throughout the parallel accounts.

We note that rigid guidelines, though necessary for strictly computational purposes, are in reality insufficient. As a contiguous series of shots, the shots in a scene are unified chiefly by strong semantics to convey a cinema story. However due to the wide berth of freedom present in both the story plot and the style with which it is told, there will always be ambiguity in the boundaries that constitute the scenes. Therefore gathering the one common thread from the aforementioned discussion, we will add one final guideline.

Guideline 5: For borderline cases arising from the application of the previous four guidelines, a strong continuity in the mood/semantics/characters/director's intent signifies the absence of a scene boundary, and vice versa.

2.4 Background and Fundamental Issues

Movie affective classification draws upon methodologies from two fields: cinematography and psychology. This section starts off by briefly introducing the necessary foundation of these two fields and the motivation for using them. We also explore various fundamental issues implicit in our approach.

2.4.1 Cinematographic Perspective

A film is made up of various elements such as editing, sound, mise-en-scene, and narrative. Governing the relationships amongst these elements is a set of informal rules known as film grammar, defined in [14] as "the product of experimentation, an accumulation of solutions found by everyday practice of the craft, and results from the fact that films are composed, shaped and built to convey a certain story." The value of film grammar to the present problem lies in the fact that it defines a set of conventions through which the meanings – many of which are affective – of cinematic techniques employed by a director can be inferred.

A quintessential example is that the excitement level of a scene increases as the shot length decreases. Other examples include rules about screen movements, cutting on action, colors and variation of lighting effects etc. By exploiting the constraints afforded by the film grammar, high level affective meaning can emerge from low level features such as shot length directly, thus offering a computable approach in bridging the difficult transition to high level semantics such as emotions. Many cues in Chapter 3 are founded on the basis of film grammar.

2.4.2 Psychology Perspective

Film evokes a wide range of emotions. Hence, a fundamental challenge of movie affective classification lies in the choice of appropriate output emotion representation in film. How do we represent emotions in movies, or relate them to existing emotion studies? These questions mirror some of the most important topics investigated in psychology, which provides emotion paradigms helpful for us in proposing reasonable answers to the questions.

A survey of contemporary theory and research on emotion psychology reveals the most dominant and relevant general theoretical perspectives, respectively known as the Darwinian [38] and cognitive perspectives [39]. The Darwinian perspective postulates that basic emotions are evolved phenomena that confer important survival functions to humans as a species, strongly implying the biological origins and universality of certain human emotions. An impressive body of evidence in human facial expression study by Ekman [16] has identified perhaps the most supported set of proposed basic emotions: Happy, Surprise, Anger, Sad, Fear and Disgust. This set of emotions, which we call "Ekman's List", are found to be universal among humans, and significantly governs our choice of output emotions and its representation.

On the other hand, the cognitive perspective postulates that appraisal, a thought process that evaluates the desirability of circumstances, ultimately gives rise to emotion. Using a dimensional approach to describe emotions under such a paradigm, several sets of primitive appraisal components thought to be suitable as the axes of the emotional space have been proposed [15], so that all emotions can be represented as points in that space. Such a representation is suited for laying out the emotions graphically for deeper analysis. The most popular appraisal axes VAD, proposed by

Osgood et al. [17] and also Mehrabian and Russell [18], are shown to capture the largest emotion variances, and comprise of Valence (pleasure), Arousal (agitation) and Dominance (control). For this work, we have found a simplified form, the VA space, helpful in visualizing the location, extent and relationships between emotion categories. Dominance is dropped because it is the least understood [33], and its emotional variance accounts for only half that of Valence and Arousal.

Outside psychology, [32] utilized a different set of emotions for machine emotional intelligence. However that set was chosen for human-computer interaction purposes, and is not suitable for describing affective content in movies.

2.4.3 Some Fundamental Issues

We first address a few fundamental issues, beginning from the emotion ground truth labeling stage: should the film affective content be evaluated according to the emotion response of the viewer or what the director intends the viewer to feel? The answer partly hinges on the nature of the currently conceived affective applications. Since they are certainly viewer centric, it is more meaningful to use viewers to calibrate the affective content. This is also consistent with the requirements of future possibilities involving personalized affective applications, which will need viewer emotion response. Not to mention that polling for directors' intentions rather than viewers' emotion responses for numerous movie scenes is far more difficult.

But this raises the question of how the inherent subjectivity of viewer emotion response should be dealt with. Some elements of uncertainty and subjectivity, depending on the unique emotion "makeup" of each individual, are inevitable in the viewer's movie experience. However the collective mean, or normative emotion

response of a statistically large audience is stable and reproducible, especially when dealing with conventional films with a body of accepted "subjective" practices and principles, and thus can be considered objective. Similar assumptions underline the validity of feedback-based psychological studies [18]. For our work, we have thus obtained this normative emotion response to movie scenes in our video corpus from a group of dedicated test subjects.

We emphasize that, though normative emotion response and director intentions broadly concur, they are not equivalent. This is apparent from the difficulties which even highly successful directors have met in conveying their visions. To us, this implies that viewer feedback is an essential element of any viewer-centric film affective system. However from the standpoint of future works involving personalized affective applications, a potential drawback is the large amount of emotion responses to scenes (of the order of a thousand) required to reliably characterize the unique emotion makeup of an individual viewer, which is too cumbersome for an ordinary user to provide. However this problem can, we feel, be greatly alleviated by casting the problem of characterizing a viewer as finding the moderately small differences between the individual viewer and normative emotion responses.

Finally, legitimate concerns on the portability of this work to movies originating from non-western cultures may be raised, given our current focus on Hollywood movies, a product of western-oriented film grammar and perspectives. For practical reasons, it is preferable to start with more established video corpus when exploring the largely uncharted territory of automated affective understanding for movies. This work does not claim universal application over movies of all origins and types. However as can be seen later, a significant portion of this work deals with

emotion features and paradigms with an underlying psycho-physiological basis common to humankind. Therefore there is reason to be confident that the work, with some culture-specific adjustments, can be validly adapted to non-western movies.

2.4.4 System Overview

We now give a system overview of our affective scene classification system. For consistency, the input to the system comprises of movie scenes manually segmented according to the criteria adopted in Chapter 2.3. For each scene, the audio and the visual signal are processed separately. The visual signal is segmented into shots and key-frames to facilitate computing visual cues for each scene. The audio

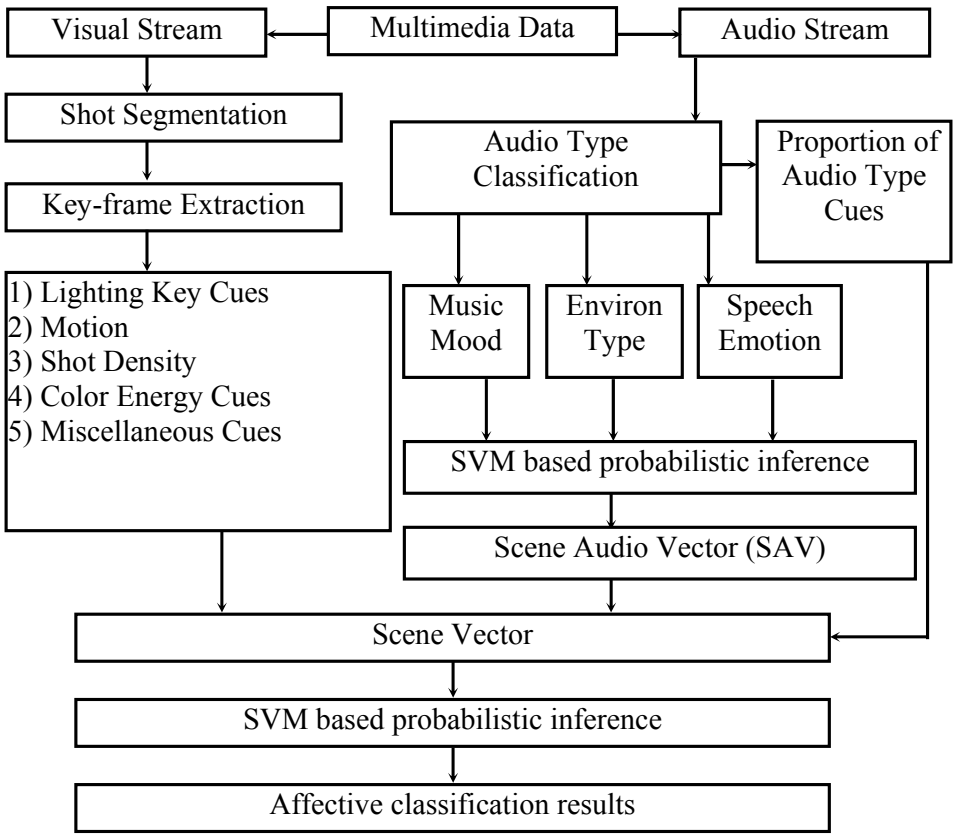


Figure 2.2 Flowchart of system overview.

signal is then separated according to audio type (music, speech, environ or silence) before being sent into an SVM (Support Vector Machines) based probabilistic inference machine to obtain high level audio cues at the scene level. The audio and visual cues are finally concatenated to form the scene vectors, which are sent into the same inference machine to obtain probabilistic membership vectors. Figure 2.2 illustrates the system overview.

2.5 Overall Framework

As a result of the intended domain of applications and perhaps to simplify matters, all prior related works have relied heavily on just one of three perspectives: Darwinian, cognitive (VA) or cinematographic. We argue for the advantages of utilizing all three perspectives in affective classification in film domain, and propose a complementary approach that for the first time exploits the information and emotion paradigms methodically from these perspectives to decide on the choice of output emotion categories and low-level input features.

2.5.1 Characteristics of Each Perspective

The cinematographic perspective provides the advantage of direct insight into film domain production rules, and is eminently suited for formulating new input features. However, its paradigm classifies film according to genre, rather than emotions. Genre is too coarse for emotion categorization, e.g. genres such as drama and romance contain a multiplicity of emotions. Nevertheless it is possible to use genre to indirectly gauge the relevance of any proposed emotion categories. The Darwinian

perspective provides the theoretical basis to categorize emotions meaningfully, but says nothing about other rich information residing in the film domain.

The cognitive (VA) perspective has the advantage of decomposing emotions into its constituent elements. Such representation offers the possibility of visualizing the entire emotion spectrum at a glance in a 2D feature space, thereby facilitating the analysis of the membership coverage and neighbor relations of different emotion categories. Due to its seeming simplicity, some works have suggested feature-to-VA mapping. But such a proposition is fraught with severe difficulties, especially when applied to the film affective domain. As further explained in the feature selection Chapter 2.5.6, this is primarily due to the complex distribution of features with respect to emotions.

However the main reason why we do not adopt the VA as the sole feature space for representing emotions is because some of the output emotions cannot even be sufficiently differentiated therein. In Figure 2.3, we graphically represented the "VA emotion space" occupied by various emotion words as ellipses, and observed that considerable overlap exists between the VA emotion spaces of emotion words associated with the basic emotions of Anger, Surprise and Fear. This overlap is confirmed by the dichotomized VA representations of output emotions (Table 2.2, 3rd column), respectively sourced from the strongest proponents of VA [18][19]. By their own accounts, VA space reveals severe to near total overlap between some output emotion categories in the VA space: namely the (Anger, Surprise), (Fear, Anger) and (Disgust, Fear) pairs. These conclusions are aligned with leading emotion theorists who criticized VA for being insufficient to "capture the differences among emotions" [36] and having "little explanatory value, and not much predictive power" [37].

2.5.2 Complementary Approach

From the strengths and limitations of being restricted to just one perspective, it is clear that affective understanding in film can benefit from a complementary approach where each perspective offers its tools and paradigms to address facets of the affective problem that it handles well and others are not able to. For our approach to retain the original theoretical bases of these perspectives, we utilize their tools and paradigms in the manner consistent with their purported theoretical strengths and properties, as examined previously and summarized in Table 2.1.

Due to a loose underlying consistency amongst the perspectives, features motivated primarily by one perspective may exhibit discernible relationships with others. Thus some may, in attempting to "unify" matters, force features arising naturally from all perspectives (e.g. the underlying physiological basis of speech audio features causes it to map more naturally in the Darwinian perspective) to map to the VA representation before mapping to the output emotions. However this hierarchical approach introduces information loss, stability and efficiency issues, especially in a complex domain such as affective classification. Instead the complementary approach fuses these features in a heterarchical manner by expressing them directly in a high dimensional space (concatenated into vectors), from which meaningful patterns can be extracted by a powerful inference engine. In this way, complex dependencies amongst features and output emotions can be captured directly, thus ensuring greater classification accuracy. The rest of the chapter will apply this complementary approach to show how the perspectives work together to decide on the choice of output emotions and low-level input features.

TABLE 2.1
Summary of Complementary Approach

Perspective	Cinematographic	Darwinian	Cognitive (VA)
Area of Strength	Most related to production of film	Organize emotions into families	Represent emotions in VA space well
Tools provided	Film Grammar	Basic Emotions	VA Space
Main Contribution	Input features	Guide to choose initial emotion categories. Input features.	Visualization of emotion membership and neighbor relationship. Input features.

2.5.3 Output Emotion Categories

A well chosen set of output emotions, besides simplifying complexity, is vital for consistent and principled manual ground truth labeling. In our view, this set should obey the following four criteria: 1) Universality: Each emotion can be universally comprehended and experienced. 2) Distinctiveness: Each emotion is clearly distinguishable from the other. 3) Utility: Each emotion should have significant relevance in the film context and finally 4) Comprehensiveness: The emotions in the set should be adequate to describe nearly all emotions in film. The first two criteria pertain to any general emotion categorization, whereas the last two are relevant for film domain application.

As discussed in Chapter 2.5.1, the cinematographic perspective offers the notion of genre as possible output emotions but the genre is a movie (and not affective) descriptor, and hence is too blunt and inappropriate for describing scene-level affective content. The Darwinian perspective offers Ekman's List which has been proven

through substantial experimental backing to be universally identifiable and distinguishable across cultural borders [40]. Ekman's List has the chief advantages of fulfilling the first two criteria, and is thus used as the principal guide in choosing output emotions.

We now want to adapt Ekman's List to satisfy the other two criteria specific to film. To begin, we investigate the relevance of Ekman's List to cinema viewers. We carried out a survey where nine respondents, each randomly assigned two movies, were asked to propose a word for each individual movie scene that would suitably describe their feelings about it, given a list of 151 emotion words found in [18] as a non-exhaustive guide. We note that because the respondents felt there were many emotionally neutral scenes, we add to the basic emotions in Ekman's List a Neutral emotion category (no emotion). The second column of Table 2.2 lists the most commonly suggested emotion words and their correspondences with the emotion categories. From the table, we can see how the more specific emotion words are related to the emotion categories. This, with one exception, attests to the utility of the basic emotions, in the sense that they can be readily associated with the more specific emotion words.

Unsurprisingly, the exception "Disgust" cannot find any correspondence with the genres and viewer feelings. This is primarily due to the lack of scenes that seek to evoke "pure" disgust in the viewer. Furthermore, cinematic scenes with an element of Disgust often contain a strong element of Fear and are thus subsumed under it. Due to its lack of utility in the cinema context, Disgust is henceforth dropped, leaving us with a set of what we term the six output emotions (Happy, Surprise, Anger, Sad, Fear and Neutral).

For the sake of comparison, Table 2.2 also lists the correspondence between the output emotions and their approximate locations in VA space (3rd column), as well as the rough correspondence with film genres (4th column). This corroboration serves to strengthen the notion that the output emotions are relevant for describing movie affective content. Note that some of the genres such as drama and romance, which

TABLE 2.2
Descriptor Correspondence between Different Perspectives

Output Emotions (Psychological)	Feelings (Viewer)	Dichotomized VA Space	Genre (Cinema)
Anger (Aggression) Exciting, Dangerous, Aggressive, Angry	Action, Adventure	-V+A +V+A	Action
Sad	Depressed, Sad, Bad, Hopeless	-V-A	Melodrama
Fear	Scary, Fearful, Terrified	-V+A	Horror
Surprise	Surprised, Tense, Anticipation	-V+A +V+A	Suspense, Thriller
Happy	Exuberance, Joyous, Enjoyment, Happy, *Heart-Warming, Tender, Sentimental, Relaxed*	+V-A +V+A	Comedy
Disgust	-	-V+A, -V-A	-
Neutral	Neutral, Boring	(V=0)-A	-

Italicized feelings corresponding to Happy differ from the non-italicized feelings in terms of arousal, a fact that will be used later. The +,- represents the positive and negative half of the Valence(V) and Arousal(A) axes.

reflect a multiplicity of emotions, cannot be matched uniquely to the output emotions and are thus omitted from Table 2.2.

To check for comprehensiveness, we used VA space as a tool to visualize the extent of coverage of the output emotions. It serves as an approximate test for comprehensiveness, in the VA sense, by showing up any large VA areas neglected by the emotions. To visualize the output emotions, we associate them to the closest related emotion words – drawn from viewer feedback in Table 2.2, 2nd column where possible – as found in the 151 emotion words list. This is because we view each group

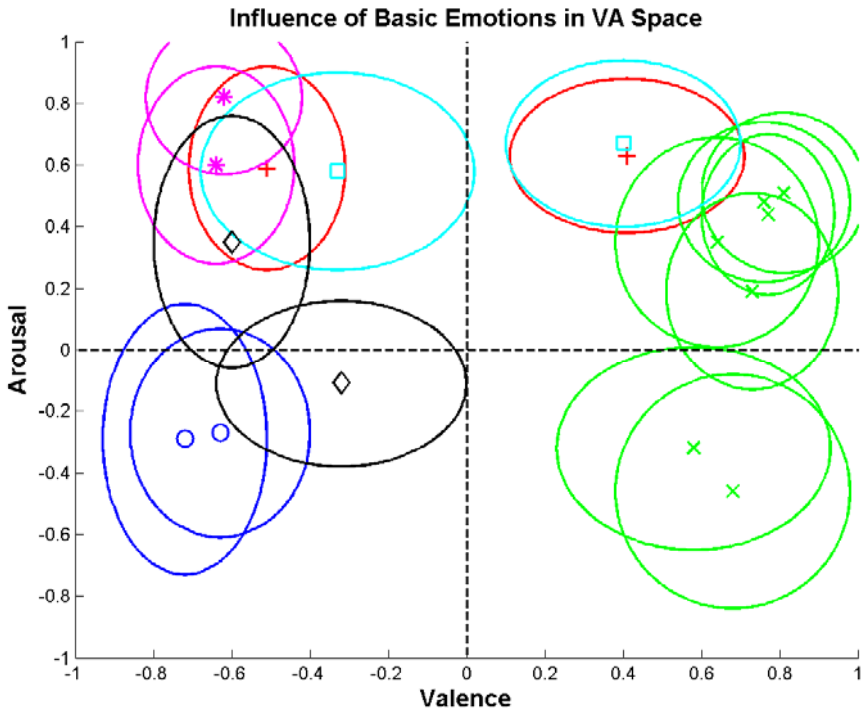


Figure 2.3 Plotting basic emotions in VA space. The VA spaces occupied by basic emotions Anger(+), Sad(o), Fear(*), Happy(x), Surprise(square) and Disgust(diamond) according to manual feedback, centered around mean and bounded by one std. deviation. The exact emotion words used are, in order of increasing arousal in each set, (Angry–anger, aggression), (Happy–relaxed, leisurely, kind, affectionate, enjoyment, joyful, happy), (Sad - sad, depressed), (Fear-fearful, terrified), (Surprise – tense, surprised) and (Disgust - disdainful, disgust).

of emotion words associated to an output emotion as “constituting a family of related affective states, which share commonalities in their expression, physiological activity, and in the types of appraisal that call them forth” [35]. These emotion words are finally mapped to the VA space in Figure 2.3, using the mean and standard deviation values of valence and arousal of those emotion words computed from manual feedback [18].

The diagram reflects the extent of coverage of the output emotions. Disgust is mapped for completeness, while Neutral, because it represents absence of emotion, does not have visualization data. The emotion areas which are more densely occupied form a rough U-shape; such distribution reflects the psychological reality that areas of high Arousal-neutral Valence and low Arousal-high Valence characterize uncommon affective states and are thus sparsely occupied. Expectedly, the only significant unoccupied region of VA space is centered on the neutral valence, low arousal area, vindicating the inclusion of the Neutral emotion in the output emotions.

2.5.4 Finer Partitioning of Output Emotions

As a natural extension of our work, we investigate into the possibility of finer meaningful output emotions. This is motivated by the fact that not all the six output emotion categories have equivalent status in the context of this work. The “Happy” emotion enjoys a privileged position in cinema, as mainstream cinema-goers still prefer to have a positive and enjoyable movie experience. This is attested by the fact that amongst the genres that can be strongly identified with an emotion (Table 2.2, 4th column), comedy is by number of movies the most popular genre (about 75,000), followed by the action genre at a distant second (about 18,000) [45].

The Darwinian perspective provides another motivation to subdivide “Happy”, which is observed to contain the most diversity in affective states [35], and hence able to yield sufficiently distinctive finer partitions. This observation is explained by the fact that four of the six basic emotions (Anger, Sad, Fear, Disgust) in Ekman’s List have immediate survival functions, and have thus evolved unique reaction patterns [36], including facial expressions, upon which Ekman’s List is based. In contrast, “Happy” contains multiple distinctive sub-families of positive feelings under it because due to the lack of immediate survival need, none have evolved its own unique facial expression away from the smiling expression [46]. This is borne out by the cognitive (VA) perspective in Fig. 2.3, which shows “Happy” affective states to be the most numerous. Hence we feel that it would be useful and cinematically relevant to extend our work by further partitioning the feelings used to describe “Happy”.

We now attempt to partition “Happy” with the four criteria of output emotion selection in mind. To retain cinematic utility, comprehensiveness and universality, we confine ourselves to defining the partitions with the feelings used by viewers to describe “Happy” emotion in Table 2.2. Referring to the italicized and non-italicized words used to describe “Happy” feelings in Table 2.2, there seems to be two sufficiently distinctive sub-families of “Happy”. The two reasons supporting this partition are as follows. Affectively, the non-italicized feelings encompass enjoyment and exuberance, and such scenes tend to be comedic or merry-making. The italicized feelings embody relaxation, kindness and tenderness, and these scenes are likely to be leisurely or heartwarming. Also in VA space, the two groups of feelings tend to be high and low in arousal respectively (Figure 2.3). Henceforth, these partitions are labeled as Joyous and Tender Affections (abbreviated as TA) respectively. As

examples, a scene where a parent comforts a child who has lost a toy falls under TA while comedic situations or boisterous friendly reunions fall under Joyous. Classification tests are carried out for both set of output emotions before and after this partitioning.

2.5.5 VA Space for Ground Truth Arbitration

To arbitrate over the output emotion assigned to emotionally ambiguous scenes that defy manual labeling of ground truth, we attempt to lay out the output emotions in the VA space (Figure 2.4) so that the entire emotional spectrum can be visualized in a glance and used as a guide. Note that Figure 2.4 is meant to conceptually depict the neighboring relationships between categories and approximate spheres of membership rather than literally demarcating crisp boundaries. In cases of ambiguity, the ground truth labeler can use the VA map as a last resort to arrive at a more objective final label.

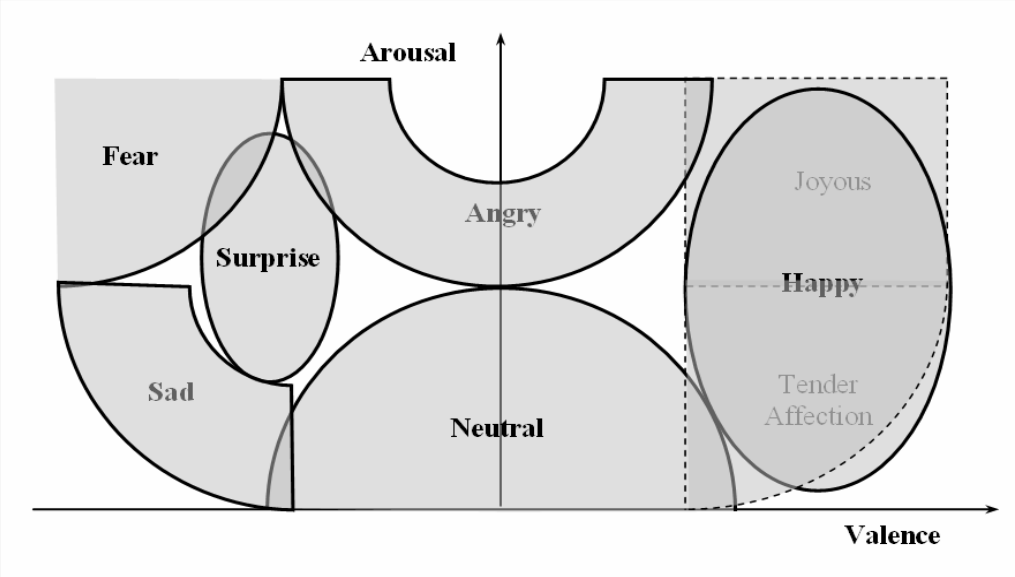


Figure 2.4 Conceptual illustration of the approximate areas where the final affective output categories (in bold) occupy in VA.

For example, in several scenes of *The Sixth Sense*, the protagonists conversed in worried tones. Although it is not immediately apparent which category "worry" belongs to, thinking in terms of where "worry" falls in VA space suggests that it should occupy the low arousal, low valence region, which coincides with Sad in Fig. 2.4. Therefore these "worry" scenes are categorized under Sad.

To function as an arbitrating tool, Figure 2.4 should avoid the emotion overlap illustrated by Figure 2.3, while preserving cinematic relevance and comprehensive coverage in the VA space. Hence it is necessarily a modification of Figure 2.3, and the boundaries of the output emotions have been suitably modified. Surprise is situated in negative valence regions to reflect the fact that Surprise scenes are mostly tense and suspenseful, as opposed to being pleasantly surprising. Aggression has also been shifted closer to neutral valence to acknowledge that such scenes are usually meant to excite, and not to provoke extreme infuriation.

2.5.6 Feature Selection

Though features exhibiting law-like or rigorous relationships with the emotion representations of various perspectives are desirable from a classification standpoint, it is realized such rigidity is incongruous with the vast artistic freedom in the film domain. For instance, just because Fear scenes tend to be dark does not imply TA scenes cannot be similarly lit. Therefore, we prefer to obtain features by utilizing guidelines stating how each perspective suggests features with significant affective implications. These guidelines are provided below, while the resultant features are detailed in the next chapter.

A) Cinematographic perspective: There are often film grammar rules with affective implications. Examples include the shot duration and lighting key introduced in Chapter 3.2. The corresponding features can then be computed. Sometimes, these rules suggest features which are mapped to VA space instead of directly to the output emotions. An example is the shot duration in Chapter 3.2.

B) Darwinian perspective: Many results in categorical perception show that the representation of some entity is critically related to its categorical membership [44], which implies that the underlying representations are different resulting from different categorical membership. The Darwinian perspective, by virtue of furnishing much of the basic emotional categories, clearly influences what features are to be used to represent these emotion categories. For instance, the audio features in Chapter 3.1, namely the Audio Type Proportion (ATP) and Scene Affective Vector (SAV), are selected (with the aid of cinematographic rules) in such a way that best separates the set of proposed emotion categories.

C) Cognitive (VA) perspective: The requirement of VA to stringently reduce features to at most two dimensions renders many complex, multimodal but informative features unsuitable for direct feature-to-VA mapping. This is especially so when features are supposed to model the high level film domain, thus limiting the usefulness of VA for feature selection. But occasionally, the concept of affective dimensions (in our case valence and arousal) recommends features with a strong connection to those dimensions. Good examples of such features are color energy and visual excitement in Chapter 3.2.

CHAPTER III

FEATURES AND EXPERIMENTAL RESULTS

In this chapter, the design and extraction of the low-level audiovisual cues used for affective classification are elaborated upon. We describe the probabilistic inference engine used in this work. Finally the experimental results are presented, followed by the conclusion.

3.1 Audio Features

Hitherto under-exploited, audio cues play an important role in this work. Five channels of information in film have been identified by Metz [13], which are: 1) the visual image, 2) print and other graphics 3) music, 4) speech and finally 5) sound or environmental effects. Interestingly, the majority of them (MSE – Music, Speech and Environ) are auditory rather than visual, implying that the auditory stream is a potentially rich source of information.

The influence of Music, Speech and Environ (MSE) audio on the affective experience has come a long way since the bygone silent film era, and its importance can be testified by the money spent in the technology and talent to create aurally realistic/powerful experiences. A main reason is because it is easier to create audio stimuli that enjoy a closer correlation with the affective content compared to the visual stimuli. This is usually due to plot, budget or venue requirements that constrain the visual composition of a scene from optimally bringing out the mood the director

intends to convey. For example, whereas the world environment and architecture in a fantasy movie like *The Lord of the Rings* may be freely designed to convey certain feelings in a visually spectacular fashion, the scope for such techniques is somewhat limited in films like *Love Actually*, which takes place in an ordinary urban setting. However the opposite is true of the accompanying music, which has much greater flexibility than its visual counterpart, as can be seen by the way music has been used in both the above movies to arouse emotions to great effect. Thus in the following audio sub-sections, we show the effective low level audio features derived based on considerations (particularly in relation to the seven output emotions chosen) discussed in Chapter 2.

3.1.1 Audio Type Proportion

Audio type classification refers to the classification of a short audio interval into one of these four types: music, speech, environ or silence. The audio type proportion (ATP) refers to the relative durations of the respective audio types to the scene duration in any particular scene. There are instances when the sound track will contain a mixture of different MSE simultaneously, however this is confusing and consequently rarely occur. The exception is speech-music, where the voice follows the affect and tune of the music (like in songs). This combination is classified as music, which is adequate for our classification purposes.

While the audio aspect of film grammar is not as formalized as its visual counterpart, we conjecture scenes with different emotions tend to possess unique ATP signatures. Both cinematographic and Darwinian perspectives suggest that different audio types are naturally suited to provoke different emotions. For instance, silence is

regularly prescribed as a tried-and-tested means to provoke surprise, whereas an aggressive scene is usually accompanied by a liberal amount of environ audio type (e.g. crashing and gunfire sounds) generated by the violence on hand.

To obtain the ATP, the entire audio stream is first divided into scenes based on manual scene segmentation. Starting and ending with the scene boundaries, type classification is carried out for every two second segment. Silence segments are first identified by thresholding the average segment energy. Two features, which are chroma difference and Low Short Time Energy Ratio (LSTER) [24], are subsequently extracted from each remaining segment. From experiments, LSTER is very effective for separating speech from music and environ segments. Finally, to differentiate between music and environ segments, we propose a simple, novel and effective feature we term “chroma difference”.

Western music is based on a reference set of 12 semi-tone musical frequencies, grouped together to form an octave. Higher and lower musical frequencies are derived by halving or multiplying all the semitones in this reference octave by a factor of two to form new octaves. Two of the properties of western music are: (a) semitones spaced an octave apart are harmonic and thus sound similar (b) sound energy is confined overwhelmingly to the semitone (music) frequencies. The chroma [28] vector, a 12x1 vector denoted as **Chr**, exploits property (a) to sum the energies for each of the 12 semi-tones across all octaves to summarize music characteristics. We exploit property (b) to differentiate music and environ with this chroma difference feature computed as

$$\mathbf{Chr} = \sum_{i=1}^{11} \left| \mathbf{Chr}_{i+1} - \mathbf{Chr}_{i} \right| \tag{3.1}$$

where $\mathbf{Chr}_i$ is the ith entry of **Chr**. Chroma difference tends to be higher in music compared to environ segments, because for environ segments, energy is usually randomly and uniformly distributed in the chroma; whereas music has high and uneven energy concentration in certain chroma bins. Segments are finally classified into its MSE type with the two features using a simple SVM.

To quantify and automatically sift out the broad patterns presented by ATP across emotions, we construct four ATP histograms – corresponding to the four audio types – for each of the seven emotions. Each of these histograms is divided into 10 equal bins [0-10%, 11-20%, 21-30%, 31-40%, 41-50%, 51-60%, 61-70%, 71-80%, 81-90%, 91-100%]. For each emotion, the bin denotes the percentage of scenes, whose duration of a particular audio type (relative to the scene) falls into the range of that bin. For instance, the 2nd bin (11-20%) of a Sad-music ATP histogram is incremented for each Sad scene with a sound track containing 11-20% of music by duration. The 28 histograms constructed are presented in Figures 3.1-3.4.

Clustering is a useful analytic tool to discover underlying patterns to the ATP histograms. However to employ clustering, some sort of distance or similarity measure needs to be defined between histograms. We adopt the use of an "Earth-Mover" Distance (EMD), which for two histograms identical in everything but their bin values, is the minimal "movement" cost required to "move" probabilities across the bins such that one histogram is transformed into the other. As an illustration, let p be the "amount of probability" shifted across $dist_{bin}$ number of bins, then movement cost incurred is computed as

$$p \text{ (probability)} * dist_{bin} \text{ (distance traversed in bin units)} \qquad (3.2)$$

Computing EMD using the brute force approach is a troublesome matter, however the following is an elegant alternative. First we compute the cumulative histogram, denoted as $\mathbf{CH}_i$, for ATP histogram $\mathbf{H}_i$. A histogram intensity measure HI_i is then computed for each $\mathbf{CH}_i$ as:

$$HI_i = \sum_{b=1}^{10} \mathbf{CH}_i(b) \tag{3.3}$$

where b is the bin index of $\mathbf{CH}_i$ and $\mathbf{CH}_i(b)$ is the value of the bth bin. HI_i tends to be higher for histograms with probabilities skewed towards the lower bins and lower for histograms with probabilities skewed towards the higher bins. Then the EMD between two ATP histograms $\mathbf{H}_m$ and $\mathbf{H}_n$ is simply

$$EMD(\mathbf{H}_m, \mathbf{H}_n) = \sum_{b=1}^{10} \left| \mathbf{CH}_m(b) - \mathbf{CH}_n(b) \right|. \tag{3.4}$$

Now, we search for patterns exhibited by the ATP histograms within each of the four audio types according to the procedure below:

Within each audio-type group (which consists of seven ATP histograms – one for each emotion)

1) Sort ATP histograms $\mathbf{H}_i$ with i=1,2,…,7 according to HI_i.

2) Compute EMD between every consecutive $\mathbf{H}_i$.

3) Split the sorted histograms into clusters between every pair of histograms m and n with $EMD(\mathbf{H}_m, \mathbf{H}_n)$>1.

4) Within each existing histogram cluster, if the EMD(first CH, last CH)>1, then split at the location of maximal EMD within the cluster to form two clusters.

5) Check that every existing histogram cluster has *EMD*(first CH, last CH)<1. If yes, stop. If not then proceed to step 4.

For our experiments, exactly two clusters are formed for each audio type histogram series. The results of this ATP clustering are shown in Table 3.1.

TABLE 3.1
Relative Audio Type Proportions For Basic Emotions

Emotions	Anger	Sad	Fear	Joyous	Surprise	T.A.	Neutral
Music	+	+	+	-	+	+	-
Speech	-	-	-	+	-	-	+
Environ	+	-	+	-	-	-	-
Silence	-	+	-	-	+	+	-

The symbols [+, -] represent the relatively low and high *HI* of the ATP histograms respectively for each emotion within each audio type group.

The cinematographic and perhaps Darwinian perspective provides some intuition to interpreting Table 3.1 and Figures 3.1-3.4. Speech is relatively heavily employed by Joy and Neutral scenes (Table 3.1 and Figure 3.1), which is not surprisingly given that most Neutral scenes are dialogue scenes. Also, a majority of Joyous scenes consist of friendly gatherings or comedic situations, where dialogue is critical to conveying the intended atmosphere or comedy. On the other hand, the rest of the emotions exhibit significantly less verbosity, particularly for Sad and Fear scenes. This is reasonable, given the reticence of sad characters (who usually appear in Sad scenes). Frightened characters (usually appearing in Fear scenes), besides being possibly reticent, often find themselves in lonely situations with no one to converse.

The patterns of environmental noise distribution across the different emotional classes show up two very distinct clusters (Table 3.1 and Figure 3.2). The first cluster consists of the Anger and Fear scenes, where environmental noise is relatively prominent. This noise typifies the loud chaotic surroundings, gunfire and noise from

frantic activities so effective in creating the atmosphere of violence and terror common to these scenes. Environmental noise plays a more subdued role for the rest of the emotions in the second cluster and is consequently minimized.

Silence frequently slips under the consciousness of the audience despite playing a subtle and important role in shaping the emotional perception. Once again, we observe two distinct clusters of silence histograms (Table 3.1 and Figure 3.3). The first cluster, which includes Sad, Surprise and Tender scenes, uses relatively liberal amounts of silence, though the mechanism whereby silence provokes the respective emotions is different. Silence enhances the soberness of Sad scenes while allowing the spectator an opportunity to engage in reflection or despair. It provides the pause necessary for engendering a sense of contentment, relaxation and quiet happiness for Tender scenes. Finally for Surprise scenes, it acts as the unsettling preamble to build audience suspense to an unbearable climax. The remaining emotions in the second cluster do not resort as readily to silence to convey themselves.

Finally, as a powerful and versatile mood inducing medium, music can be used to provoke almost any kind of emotion, as can be seen from the nearly uniformly distributed music proportion histograms (Table 3.1 and Figure 3.4) for the majority of categories. In fact, although one expects melancholic music to play a major role in Sad scenes, it is still rather remarkable to note that the sound tracks from 10% of all Sad scenes are pure music compositions. The only emotion classes that under-weigh the use of music would be the Neutral and Joy categories. Due to the lack of emotional content, Neutral scenes have little need for music.

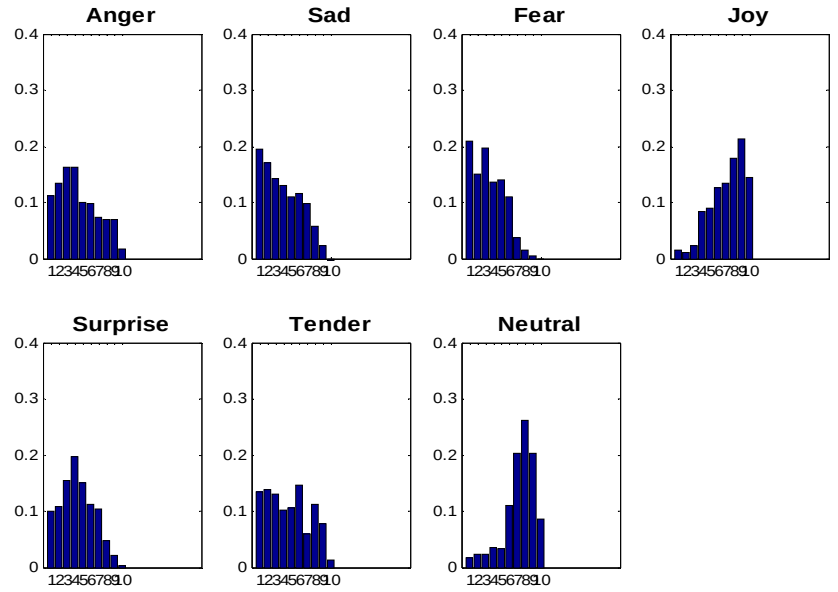


Figure 3.1 Speech audio proportion histograms for emotional classes.

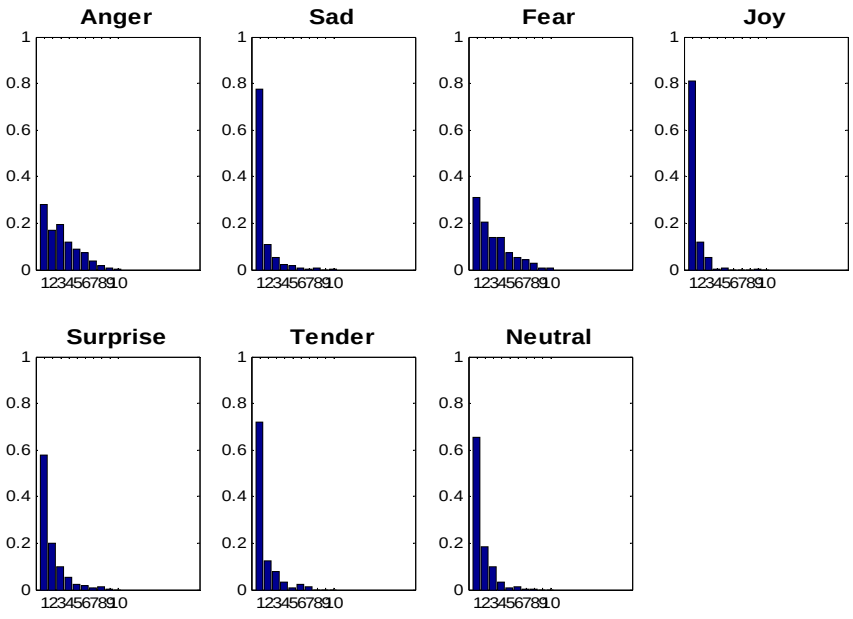


Figure 3.2 Environ audio proportion histograms for emotional classes.

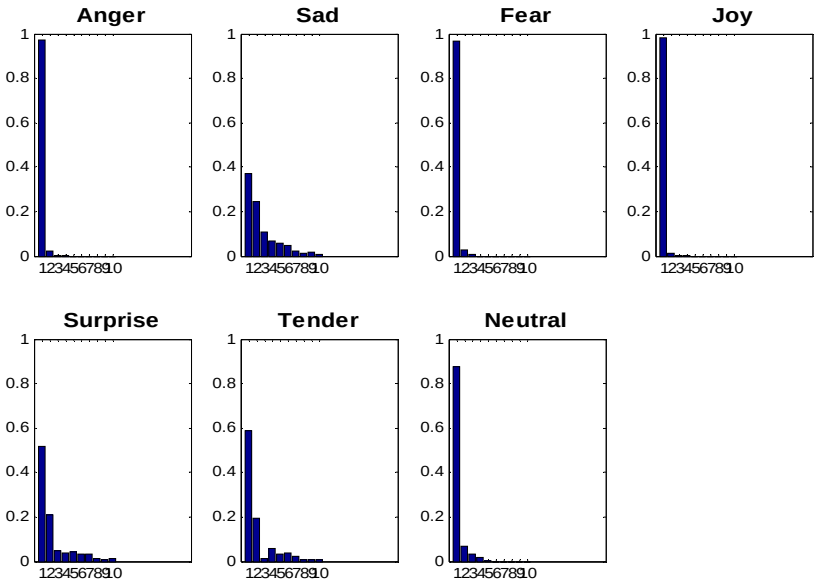


Figure 3.3 Silence audio proportion histograms for emotional classes.

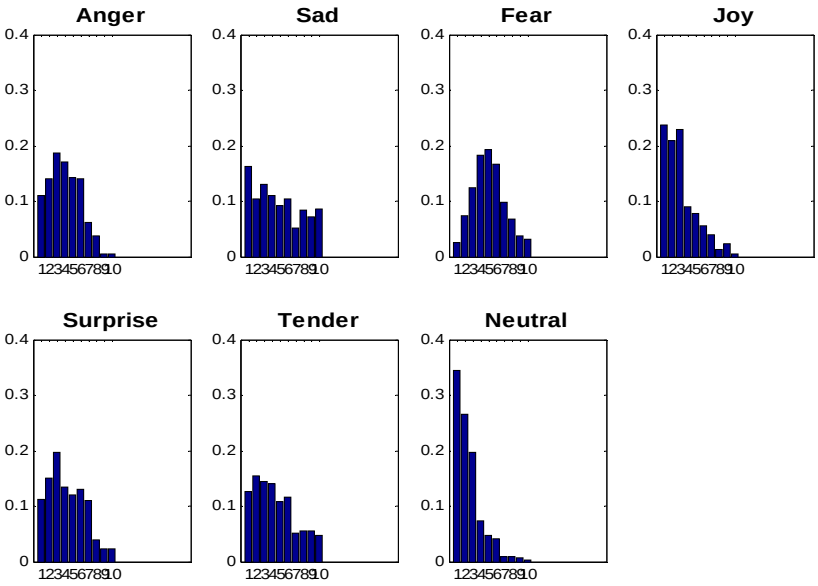


Figure 3.4 Music audio proportion histograms for emotional classes.

3.1.2 Audio Scene Affect Vector (SAV)

As pointed out, each MSE channel contains significant affective information. Owing to their origins and as well as semantic meanings, these three indispensable audio components of modern cinema are significantly different and in turn have to be treated in their own appropriate and unique ways. Amongst the triad, speech is the most indispensable. This is due to the predominance of the usage of speech in life, which necessitates its screen portrayal. Serving mostly as the primary vehicle for humans to interact one with another and to advance the plot, it also informs the spectator via devices like narratives and voice-overs. Due to the predominance of speech, and strong linkage between speech features and output emotions, speech plays the most important role amongst the triad. Music is similarly informative but due to its modest usage is slightly less influential. Lastly, though environ sound is the least distinctive of the triad, it can still distinguish between broad sets of emotions.

This suggests the approach of computing appropriate low level audio features that contain affective information for each MSE type, and combining this information across a scene to obtain the scene affective composition, also known as the audio Scene Affect Vector (SAV). The SAV is a vector denoting the amounts, or probabilities, of the output emotions existing in a scene based on the audio track. Since the dimension and nature of the SAV are solely dependent on the exact output emotions chosen, it is intimately related to the Darwinian perspective.

For a brief review, high accuracy for speech and music mood detection has been achieved by the current state of the art. For instance, New [26] achieved 78.1% accuracy in classifying speech emotions for Ekman's List of emotions. The dataset comes from twelve speakers, reading standard scripts under laboratory conditions. Liu

[25] obtained 86.3% accuracy in classifying music clips into the four classes of contentment, exuberant, anxious and depression, with data consisting of meticulously prepared clips from the classical and romantic period.

In contrast to the specially prepared data of those works, the vast diversity of movies has thrown up for us a very difficult dataset. The speech segments alone contain speech spoken by diverse races, gender, age groups and in multiple English dialects, styles, pitch, speed and volume. Similarly, the music segments feature large diversity of styles from different eras, generated from different instruments. In addition, the movie soundtrack is often interspersed with significant amount of incidental noise. It is under these trying circumstances that music and speech are both classified into seven output emotions.

In order to effectively integrate the audio information from all the MSE types at the scene level, we have proposed the following algorithm (Figure 3.5). Initially, audio segments belonging to the same MSE type and scene are concatenated together, subject to only one constraint: speech segments cannot concatenate across shot boundaries, because they usually denote speaker change, whereas music easily stretches across shots. These concatenations are then partitioned into affect units of 8, 4 and 2 seconds duration for music, speech and environ sound respectively. These durations are chosen because they are typically the least time required for a confident assessment of the affective content for the corresponding MSE type. For instance, four seconds is about the least time required to aurally discern the emotion of a speech fragment. Affect units are then labeled with the output emotion of the scenes they belong to. Each MSE type requires a different set of suitable features [24]-[28] to recover its affective content. Thus for each affect unit, we compute the set of features

appropriate for its MSE type, and organize these features into a vector. Here we present the details of the features and the basis for using them.

Automated classification of speech emotion has been a topic of research for about fifty years: it has been discovered that psycho-physiological characteristics like air intake, vocal muscle, intonation and pitch characteristics vary with emotions. Computable speech features, also known as prosodic features, are able to measure these characteristics, hence directly aiding speech emotion classification. To compute the prosodic features for every 4 second speech affect unit, it is divided into 40 frames each lasting 100 milli-seconds. Discrete Short Time Fourier Transform is then performed on the signal of each frame to calculate the following features:

Spectral shape features [25][42]: a) centroid, b) bandwidth, c) roll-off and d) spectral flux. These four features provide broad information on the shape and range of the spectral energy distribution for each frame, and how fast it varies from frame to frame, as emotion affects the energy distribution of voiced speech.

Raw power feature: power of the unprocessed audio signal is used as a feature.

Log Frequency Power Coefficients (LFPC) features: The most notable of prosodic features is a set of 12 frequency sub-bands [26], across which the distribution pattern of spectral energy is able to distinguish between different emotions in human speech. The energies of these twelve sub-bands are known as the Log Frequency Power Coefficients (LFPC), and constitute the key prosodic features used.

Since the abovementioned 17 features are extracted from each of the 40 frames belonging to a speech affect unit, it is possible to calculate both the mean and variances for each individual feature above with respect to the affect unit. This results in a 34x1 vector –17 from the means and 17 from the variances of each feature – to describe a speech vector.

For music mood identification, there is little consensus at present on the exact mechanism whereby music evokes emotions. However numerous references such as [43] and those found in [25], in accordance with established music knowledge, agree that aspects like music mode, intensity, timbre and rhythm play important roles in evoking different musical moods. To compute the music related features for a 8 second music affect unit, we have divided the unit into 40 frames each lasting 200 milli-seconds. Discrete Short Time Fourier Transform is then performed on the signal of each frame to calculate the following features:

Spectral shape features [25][42]: a) centroid, b) bandwidth, c) roll-off and d) spectral flux. These four features outline the shape and range of the spectral energy distribution for each frame, and how fast it varies from frame to frame, providing some useful information on the timbre of the music.

Raw power feature: power of the unprocessed audio signal is used as a feature.

Chroma features [28]: The energies of the frequency sub-bands belonging to the semitones in 7 consecutive octaves, from the lowest note as C1=32.7Hz to the highest note at C8=4186Hz, are extracted using 84 filter banks, due to the 12 semitones in

each of the 7 octaves. The chroma vector, a 12x1 vector, is obtained by summing the energies of the same type of semitones across the octaves.

Dominant Music Scale feature: According to the influential "Doctrine of the Affections" [20], the music scale measures the valence of a music piece (minor scale for sad and major scale for happy). Both minor and major scales have their own unique semitone interval patterns which restricts the set of permissible semitones playable. Exploiting this fact, it is possible to compute the music scale of a frame: the energies of various semitones in the chroma vector are summed up according the individual sets of permissible semitones [133] for both scales, and the scale that receives the higher energy for this summing operation is deemed to be the scale of the frame.

There are in total 17 spectral shape, raw power and chroma features extracted from each of the 40 frames belonging to a music affect unit. Similar to speech, the mean and variances of these individual features are computed with respect to the music affect unit, resulting in a 34x1 vector – 17 from the means and 17 from the variances of each feature. Two additional sets of features: a) the normalized mean chroma vector of the affect unit and b) the percentages of music frames classified as major and minor scale within the affect unit are added to the 34x1 vector, bringing it up to a 48x1 vector used to describe the music affect unit.

Environ sound can originate from many different sources and corresponds only very loosely to certain sets of emotions. Consequently, there are few features that particularly lend themselves to characterizing environ affect units: the mean and variance of the audio signal are used to represent each environ affect unit.

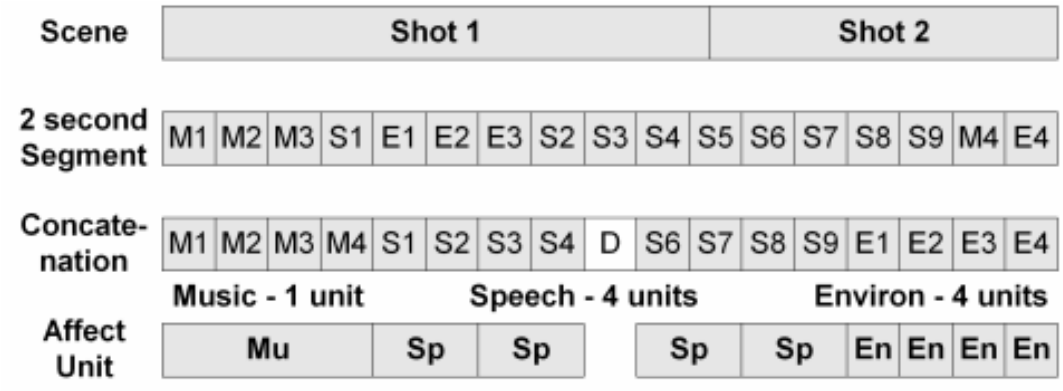


Figure 3.5 Illustration of the process of concatenating the segments into affect units to be sent into the probabilistic inference machine. D is for discarded speech segments that straddle shot boundaries.

At this stage, each affect unit is represented by a feature vector of the appropriate features for its MSE type. All feature vectors of all affect units are divided into three sets according to MSE type. For each of these three sets, the feature vectors therein are divided into K=10 groups; with each group being sent into a SVM probabilistic inference machine to obtain the output vectors V_{au} while the remaining K-*1* groups function as training data. This inference machine, which is mentioned in detail in Chapter 3.4, takes an input feature vector and outputs a $N \times 1$ row vector V_{au}, where N is the number of affective categories and the entries of V_{au} represent the probabilities of the feature vector belonging to the respective categories. Let the number of affect units, indexed by i, in a scene be N_s. Let their corresponding durations be t_i and output vectors be $V_{au,i}$, then

$$\text{audio Scene Affect Vector (SAV)} = \frac{\left(\sum_{i=1}^{Ns} t_i V_{au,i}\right)}{\left(\sum_{i=1}^{Ns} t_i\right)}. \qquad (3.5)$$

The SAV, which constitutes part of the final scene audio cues, possesses several advantageous qualities. Firstly, it is time weighted to accurately reflect the contribution of every classifying unit. Secondly, since the output vectors $V_{au,i}$ are probabilities, the SAV has a natural probabilistic interpretation; each SAV entry (SAVE) denotes the probability of the scene belonging to the corresponding category like V_{au}. Thirdly, due to the integration of information from many affect units, the SAV is far less prone to outlier errors. Finally, using affect units of short durations better models the possibility of affects changing throughout the scene.

If we use an ordinary SVM to classify individual speech and music affect units, the overall classification performances are 45% and 40% respectively. The main reasons for such dismal results are because 1) the training samples are insufficient and 2) the movie sound tracks contain many instances of MSE type mixing, which are much harder to classify than audio with one pure MSE type. For comparison purpose, we utilize the SAV probabilistic framework to construct two modified audio SAVs for each scene: one constructed using only speech affect units and the other using only music affect units. If the class receiving the highest probability in each modified audio SAV is the test label of the vector, then the speech and music modified audio SAV will enjoy classification accuracy of 65% and 57% respectively.

3.2 Visual Features

We describe several visual cues and show their relationships with respect to the perspectives laid out in Chapter 2. Here we state some important preliminaries that apply to the visual features described in this section. Unless otherwise stated, the visual cues are computed exclusively in the *HLS* (Hue, Lightness, and Saturation)

color space. This is justified purely on the psychological evidence [19] that humans perceive the "emotional" influence of colors with respect to its Hue, Lightness and Saturation components. *HLS* histograms, where applicable, are generated by dividing each axis into 20 equal intervals.

For reasons of computational efficiency, all visual features, except shot duration (Chapter 3.2.1) and visual excitement (Chapter 3.2.2), are computed from key frames. The first frame of every shot is declared a key-frame and further key-frames from each shot are then selected according to the method detailed in [2], which selects a current frame as a key-frame if its visual difference with the previous key-frame exceeds a threshold. Let the key-frame feature (KFF) for visual feature z of key-frame KF_i be KFF[z, KF_i] and t_i be the number of ordinary frames that key-frame KF_i represents. Then the scene feature SF[z, sn] of visual feature z for the whole scene sn with N_{KF} key-frames is

$$\mathrm{SF}[z, sn] = \frac{\sum_{i}^{N_{KF}} (t_i \, \mathrm{KFF}[z, KF_i])}{\sum_{i}^{N_{KF}} t_i}. \tag{3.6}$$

3.2.1 Shot Duration

From the cinematographic perspective, the perceived passage of time, also known as the pace, is manipulated to great effect by editing effects like cuts, which defines the shot length. As each shot conveys an event, the director can heighten arousal and intensify a scene by increasing the event density via rapid shot changes [22]. To the viewer, rapid shot changes capturing the main action from different angles

certainly convey the dynamic and breathtaking excitement far more effectively than a long duration shot [11][23].

Shot boundary detection is an essential first step. However since the system only needs to detect intra-scene shot boundaries, it does not need to consider challenging editing effects like dissolves or fades more often used for inter-scene boundaries. We start by computing the one dimensional shot boundary profile, **M**, which measures a certain visual distance between every frame pair F1 and F2. Let F1 and F2 be similarly tessellated into 20×20 blocks, and let an *L* histogram (as in the *HLS* color space both these frames are represented in) be constructed for each of these blocks. Then *M*(*F1,F2*), or the visual difference between F1 and F2, is defined as the sum of the L2-norm between *L* histograms of all corresponding blocks of F1 and F2. Let *j* be a general frame index, then shot boundary is declared at frame *i* only if the frame fulfils these three criteria C1(*i*), C2(*i*) and C3(*i*):

$$C1(i): \left|\frac{\partial \mathbf{M}}{\partial j}\right|_{j=i} > 0.06, \quad C2(i): \left|\frac{\partial^2 \mathbf{M}}{\partial j^2}\right|_{j=i} > 0.04$$
$$C3(i): (C1(j)) \wedge (C2(j)) \ \textit{iff} \ (j == i) \ \forall \ (i\text{-}15) < j < (i+15) \tag{3.7}$$

C1(*i*) ensures the frame-pair frame visual difference is above a minimal value shot boundaries are empirically observed to exceed. C2(*i*) is used to disambiguate a true shot boundary from consecutive high frame visual difference due to fast or large moving objects (the latter case would have small $\left|\frac{\partial^2 \mathbf{M}}{\partial j^2}\right|_{j=i}$). Finally, C3(*i*) encodes the condition that shots must last for a perceptible length of time. Average recall and precision rates are around 94%. The average shot length of a scene is then calculated as (total scene duration)/(no. of shots in scene).

3.2.2 Visual Excitement

Motion plays a central role in the cinema experience owing to the intimate correlation between the degree of mental excitement and the perception of motion on screen. This correlation, broadly proven by a psycho-physiological study [21], seems to result from the natural association of fast motion with danger and excitement, as well as new activity or information. From the cognitive (VA) perspective, computing the arousal arising from motion, which we call visual excitement, is useful in differentiating between emotions in different halves of the arousal axis (Figure 2.3). Hence we explore a method to accurately determine this visual excitement by the motion present in a video sequence.

Existing approaches [9]-[11] have proposed reasonable features to measure visual excitement. However in the affective context, those features suffer from a somewhat arbitrary mapping to visual excitement. In contrast, our proposed feature is actually obtained from a non-linear regression of actual psychophysical results obtained for visual excitement, thus reflecting the critical link between the low-level feature and visual excitement.

Our visual excitement measure is based on the average number of pixels between corresponding frames that have changed according to human perception. This change is computed in the perceptually nearly uniform *CIE Luv* space, since visual excitement is intended to model human perception, and frame difference (L2-norm) calculations are required across the entire spectrum of possible colors. To compute this visual excitement measure between two consecutive frames F1 and F2, they are both similarly tessellated into 20x20 blocks. Let (L_0,u_0,v_0) and (L_1,u_1,v_1) be the average *CIE Luv* values of corresponding blocks from F1 and F2, and let the average frame

luminance be $s_{avL.}$ To smooth over noise, the frame difference is calculated over a 20×20 block (Figure 3.6) as

$$x_{fd} = \sqrt{s_L (L_1 - L_0)^2 + 1/3\left((u_1 - u_0)^2 + (v_1 - v_0)^2\right)}$$
$$s_L = \begin{cases} 1/3 & , \ s_{avL} \geq 1/3 \\ 1/3 + \left(s_{avL} - 1/3\right)^2 & , \ otherwise \end{cases} \tag{3.8}$$

where a block is declared as changed if x_{fd} is greater than threshold $thres_{fd}$=7. Since frames low in s_{avL} tend to return lower visual excitement values, a scaling factor s_L is used to increase the sensitivity of x_{fd} to luminance differences for darker frames. Let H be the Heaviside step function, N_H the number of blocks in a frame and let k index the blocks of each frame. Then X_{fd} is defined as

$$X_{fd} = \sum_{k=1}^{N_H} H(x_{fd}(k) - thres_{fd}) / N_H \tag{3.9}$$

Finally the visual excitement for each scene is computed as

$$(1/N_c) * \sum_{f=1}^{N_c} [X_{fd} + (X_{fd})^W]_f \quad \text{s} \tag{3.10}$$

where N_c is the number of frames in the scene and f indexes the frame. In order to prevent bias towards slow motion clips, we add an offset bias $(X_{fd})^W$ where W is a constant whose optimal value is empirically determined later.

To determine the optimal parameters for the visual excitement measure as objectively as possible, a diverse test set comprising of 82 video clips of various types and degrees of motion and lighting conditions are manually selected and segmented from seven movies. These clips feature explosions, large occlusions and special effects averaging around 15 seconds each and have only one type or degree of motion.

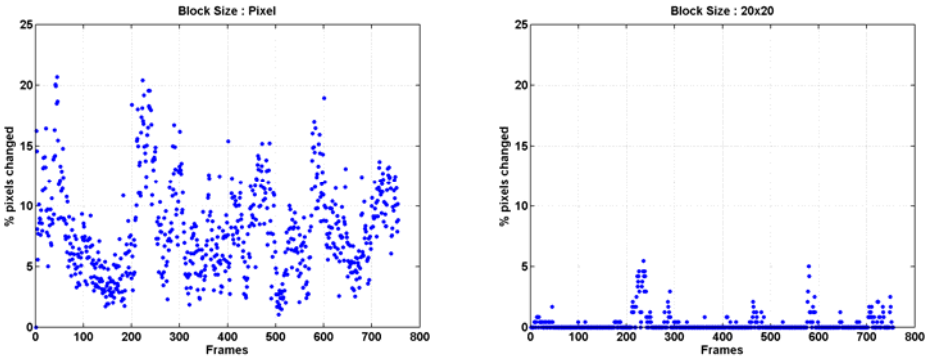


Figure 3.6 The amount of pixel change detected (%) for each pair of consecutive frames using pixel sized (left) and 20x20 blocks for a video clip. Experimentally, the plot for the 20x20 block size follows human perception of motion closely, illustrating the smoothing benefits of aggregating pixel change over blocks.

Three test subjects are instructed to give an approximate score to each clip, as far as humanly possible, according to how the motion (not the content) excites them. The clips that each test subject feel to be the most exciting and sedate are assigned ten and zero respectively, and used as reference (calibration) clips (Figure 3.7). Finally all the rest of the clips are scored manually on a linear excitation scale of zero to ten. Based on the scores, a proposed regression function with suitable parameter value of W=0.75 is obtained, with errors ranging from 0.1% to 13.45% and a mean of 3.91%. From the results (Figure 3.8), it is observed that the measure correlates very closely with the manual scale ranking. As mentioned before, few informative affective cues are suitable for such rigorous regression to even a very limited definition of psychological arousal (visual excitement). However the results show that the proposed measure is indeed an acceptable indicator of visual excitement.

Figure 3.7 Video clips of various speeds on the scale of 0-10, arranged row-wise, with slower clips at the top and faster clips at the bottom.

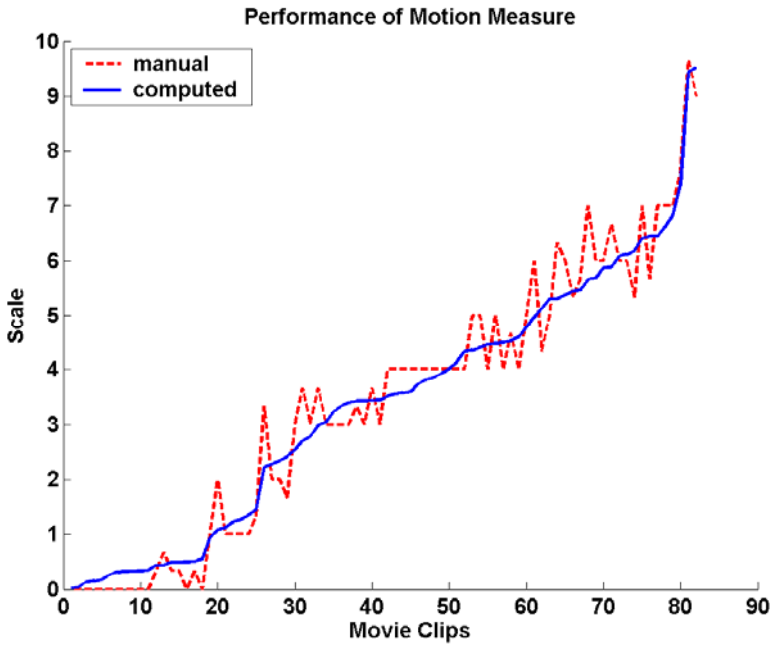


Figure 3.8 Graph of the computed visual excitement measure plotted against the manual scale ranking for each movie clip.

3.2.3 Lighting Key

In the cinematographic perspective, lighting is an extremely powerful tool, used specifically for the purpose of affecting the emotions of the viewer and establishing the mood of a scene. Generally two major aesthetic lighting techniques are frequently employed. Low-key lighting, or chiaroscuro lighting, is characterized by a contrast between light and shadow areas whereas high-key, or flat lighting, deemphasizes the light/dark contrast [22][1]. To generate the light-heartedness and warm atmosphere typical of TA and Joyous scenes, an abundance of bright illumination and a light background, in the form of high-key lighting, is usually employed. In the same vein, film grammar prescribes the use of dim lights, shadow play and predominantly dark background to recreate the Sadness, Fear and Surprise for sad, frightening or suspense scenes [22].

From the above definitions of the two lighting keys, their differences are determined by two factors: 1) the general level of light and 2) the proportion of shadow area. [11] proposed detecting lighting key using the product of the mean and variance of the brightness of a frame. However the mean is very sensitive to extreme values. The variance is also not discriminative enough because it only measures the amount of deviation from a mean, and as such a high-key lighting frame can easily have the same variance as a low-key lighting frame.

We have therefore attempted to formulate two visual features that can accurately quantify the aforementioned components of lighting key in order to better detect it. The median, Med_l, is used as an indicator of the first component, which is the general level of brightness, due to its robustness in the presence of extreme values. The second component, the proportion of shadow area, can be characterized by using the

proportion of pixels, Pro_s, whose lightness fall below a certain shadow threshold Th_s. This threshold is experimentally determined to be 0.18, at which an average saturation and highly textured surface no longer appears as textured.

3.2.4 Color Energy and Associated Cues

Psychological studies on color have shown that valence is strongly correlated to brightness and to a lesser extent saturation while arousal is strongly correlated to saturation [19]. Thus to capture these affective relationships, we have introduced what we call the color energy cue. This cue depends proportionally on the saturation, brightness and area occupied by the colors in a frame [22]. It depends also upon the hue, as in whether it contains more red (energetic) or blue (relaxing) components and the degree of contrast between the colors [22]. From the cognitive (VA) perspective, color energy measures the joint valence-arousal quality of a scene arising from the color composition alone. Thus, the degree of valence or arousal in a scene can be partially inferred by its color energy. For instance, a Joyous effect can be manufactured by setting up a scene with high color energy.

Let i, j index the bins in the *HLS* histogram of an image and p(i) denote the histogram probability for bin i. Let d(i,j) denote the L2-norm in *HLS* space between bins i and j while M is the total number of pixels, over which index k iterates while s_k, v_k and h_k are respectively the saturation, lightness and hue values of pixel k. Color Energy is defined as the product of raw energy (first term) and color contrast (second term) as:

$$\left[\sum_{k}^{M} \mathrm{E}(h_k) s_k v_k\right] \times \left[\sum_{i}\sum_{j} \mathrm{p}(i) \times \mathrm{p}(j) \times \mathrm{d}(i,j)\right]. \tag{3.11}$$

The first term attempts to approximate the arousal caused by the color composition of an image by summing the product of s_k and v_k together, both variables that supposedly correlate positively with arousal. The E(h_k) hue function weighs the resultant product with a value between [0.75-1.25], depending on the relative angular distance of h_k to blue (relaxing – lower energy) and red (energetic – higher energy) respectively. In *HLS* space, the hue values of red and blue are 0 and 240 degrees respectively. Let $dist_{red}$ and $dist_{blue}$ be the minimal absolute angular differences of h_k from 0 and 240 degrees respectively in a 360 degrees hue system (wrap-around from 360 to 0 degrees allowed), then E(h_k) is defined as

$$\mathrm{E}(h_k) = \frac{1.25 * dist_{blue} + 0.75 * dist_{red}}{dist_{blue} + dist_{red}} \tag{3.12}$$

The second term measures contrast within an image in terms of the colors and their amounts that make up the image. Colors that are very dissimilar generally produce greater contrast between themselves in the image and vice versa, which we attempt to capture with d(i,j). The function d(i,j) is multiplied by the product of p(i) and p(j) to factor in the greater influence of histogram bins with higher values.

3.3 Inter-Feature Relationships

To quantify the amount of correlation between each individual feature, we have computed the correlation coefficients according to

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \tag{3.13}$$

where *C(i,j)* is the covariance. The correlation matrix is illustrated graphically in Figure 3.9, with lighter cells denoting high correlation. From this matrix, several

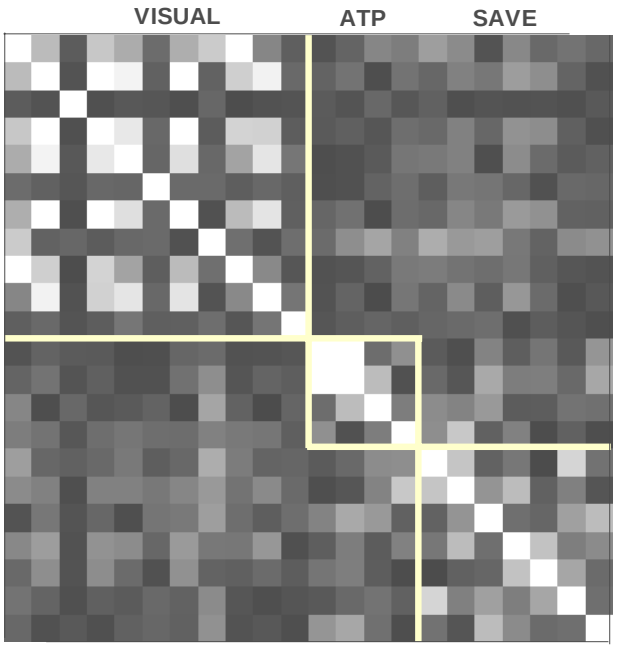


Figure 3.9 Feature correlation matrix.

observations can be made. Firstly, the level of intra-correlation of the three major groups of features is significantly stronger than the inter-correlation. At the same time, it may suggest some feature redundancy within especially the visual group. However given that some correlation is almost certain to occur within features of each group, and the small number of features relative to the power of the SVM, there is insufficient justification for feature pruning, which may harm classification accuracy.

Secondly, the lower level of inter-feature correlation suggests that using audio features in addition to visual features does enhance classification accuracy because audio features contain information not available to the visual domain. Furthermore, it is interesting to note that though both ATP and SAVE are derived from audio information, the inter-correlation between them is not strong. This is because ATP deals with duration whereas SAVE mostly deals with how energies in specific

frequency bands vary with time, confirming that the current separate grouping of these two groups is appropriate.

3.4 Classification and Inference

The affective features as described in previous sections are extracted and concatenated into row vectors to form the data points characterizing every scene. Due to the highly irregular nature of their probability densities, the classification method needs to be selected with care. In particular, no artificial constraint should be foisted on the data. This excludes parametric based methods or ad hoc rule based methods from consideration. In view of the requirements, we use a specially adapted variant of Support Vector Machines (SVM), which has proven highly successful for classification.

The SVM, which can map vectors from an input space into a possibly infinite dimensional feature space and find the best separating hyperplane therein, has only two adjustable parameters and tends to be less susceptible to the curse of dimensionality [29]. This scheme does not make unwarranted independence assumptions regarding the interaction between audio and visual cues; any such interaction is left to the SVM to learn and exploit. As a kernel based method, it is also able to model extremely complicated class boundaries. Finally, the variant used allows flexibility in the features chosen and more importantly, outputs *a posteriori* probabilities for every category, permitting more refined characterization than binary outputs and allowing the presence of multiple emotions.

To begin with, the data are normalized by shifting the centroid to the origin before dividing it by the mean of the absolute magnitudes. Then 10-fold cross

validation is used with grid search to obtain the optimal penalty and margin parameters. Subsequently, radial basis kernel SVM classifiers are individually trained for all unordered class-pair combinations. For instance, if there are seven classes in this work, then there will be (7*6/2)=21 class-pairs. In line with the ambiguous nature of those training data with dual labels, these data are included in the training sets of both classes of the class-pair SVM classifiers being trained, and excluded only if the SVM class-pair coincides with the dual labels. *A posteriori* sigmoidals fitted to the decision values of the SVMs are learnt for each class-pair [30]. The sigmoidals, with a, b as adjustable parameters, are of the form

$$p_i = \frac{1}{1+\exp(af_i + b)} \tag{3.14}$$

where p_i is an *a posteriori* and f_i a decision value. These sigmoidals are shown to be very good in modeling the *a posteriori*. The training phase ends when the parameters of all class-pair sigmoidals have been obtained.

For the testing phase, a test vector v_t – which in the general case is a vector of features representing a unit to be classified – is then processed by each class-pair SVM to obtain the decision values, which are in turn fed into the respective sigmoidals to produce class-pair probabilities. These probabilities are finally combined together to output a N_cx1 row vector where the entries are the *a posteriori* of v_t belonging to each of the N_c classes [31]. For our work, the test vector v_t is a vector of features representing either a movie genre, scene or shot depending on the specific application as seen in later sections.

3.4.1 Exploitation of Scene Temporal Relationship

A natural inquiry into the affective analysis of Hollywood multimedia is this: do temporal relationships exist at a suitable level of abstraction that enables them to play a useful role in classifying scene affective information? At first sight, the answer is apparently positive. Afterall, Kang [10] and Zhai [12] have modeled such dependencies at the shot level using HMMs and FSMs respectively, with seemingly good results. However there are some serious caveats. In both works, only three emotion categories are specially chosen for the selectors' algorithms. Furthermore, the input data are carefully selected and extremely limited at less than two hundred scenes. From our data, we observe there is a tendency for scenes of the same type to appear together. However it is not clear whether there really are significant causality relationships amongst scenes aside from this phenomenon.

To investigate this issue, we carried out an experiment to ascertain if there is any advantage to adopt a temporal generative model like HMM as opposed to a discriminative model like SVM. Assuming Markovian distribution for scene labels, the idea is to obtain the probability distributions of the forward Markovian probabilities and prior forward probabilities between different categories and conduct a test for any significant statistical differences. Let $P_M(c_i,c_j)$ denote the Markovian forward probability of scenes of class c_i preceding scenes of class c_j, and $P_a(c_i,c_j)$ denote the *a priori* probability of scenes of class c_i and c_j being adjacent to each other. If there is any exploitable temporal information, we would expect the probability distributions of $P_M(c_i, c_j)$ and $Pa(c_i, c_j)$ to be different.

Because the Kolmogorov-Smirnov test (KS-test) [130] does not make assumptions about the distribution of data, meaning it is non-parametric and

distribution free, the test is used to determine whether the distributions $P_M(c_i, c_j)$ and $Pa(c_i, c_j)$ are different. We compare the seven column-wise one dimensional PDFs of $P_M(c_i, c_j)$, one at a time, with $Pa(c_i, c_j)$, which is constant, at a statistical significance of 5%. The results show that of the seven comparisons, only the $P_a(c_i,c_j)$ of Fear is significantly different from $Pa(c_i, c_j)$. Additionally, we used this data to construct a HMM inference system with the observation likelihoods provided by the SVM probabilistic output. However the results are poorer than using the SVM inference alone.

We can only conclude that although there are certain exploitable temporal relationships amongst scenes, they are generally not strong nor widely applicable enough across all types of scenes. Secondly assumptions on scene label distribution may not even be appropriate for this problem. This argues against the use of specific temporal modeling techniques like HMM and FSM, which can possibly harm classification accuracy, although the use of other types of temporal modeling methods may yet prove beneficial.

3.5 Experimental Results

Our training data consists of 36 full-length and mostly recent mainstream Hollywood movies chosen to represent the more popular films. This translates into 2040 scenes, whose percentage distribution by output emotions are Neutral(24%), Fear(8%), Joyous(13%), Surprise(16%), TA(11%), Anger(17%) and Sad(12%). There is also a diversity of directorial styles so that the training scenes are likely to be unbiased. Table 3.2 divides these films according to the major genres.

TABLE 3.2
Movies Used For Affective Classification

Action	Horror
The Fifth Element	Ghostship
Speed	Queen of the Damned
Lord of the Rings I	The Haunting
James Bond (Golden Eye)	What Lies Beneath
True Lies	The Others
Men In Black	Dream Catcher
Saving Private Ryan	Ring
Starship Troopers	Gothika
Star Wars I	Legend of the Mummy
Waterworld	
Jumanji	
Drama/Melodrama (D/M)	**Romance/Comedy (R/C)**
Forrest Gump	There's Something About Mary
Magnolia	My Best Friend's Wedding
Ghost	Up Close and Personal
Life is beautiful	Bedazzled
City of Angels	50 First Dates
Artificial Intelligence	Maid in Manhattan
The Sixth Sense	Love Actually
	Bruce Almighty
	Notting Hill

3.5.1 Manual Scene Labeling

To obtain the ground truth for experimentation, we attempt to manually match the affective content of a scene to one of the output emotions. If ambiguities arise, we resort to the VA diagram. Three persons are assigned to independently label each scene. To prevent fatigue and systematic bias, an individual labels only one random

movie daily, of a genre different from the previously labeled movie. Except for unanimous decisions that stand, all scenes with dissenting views are reviewed using Figure 2.4 as a guide, which usually result in common agreement. Scenes where no agreement can be reached have dual labels; the main label that received two votes, and an alternate label that received one vote. Dual label scenes comprise 14.08% of all scenes; there are no cases with three differing votes.

3.5.2 Discussion

The testing is carried out by take-one-movie-out testing method using our system, whereby the scenes of an entire movie is used for testing while the remaining scenes are used for training. This method of testing, where every scene is classified into an output emotion, is repeated for every movie in Table 3.2. The results for all the movies are aggregated and presented in the form of a confusion matrix given in Table 3.3 for the extended framework. For a clearer analysis of the algorithm performance, the confusion matrix is presented (Table 3.4) in the form of the confusion rate between every pair-wise emotion. Let $\mathbf{c}(i)$ denote the set of scenes in emotion class i, $|\mathbf{c}(i)|$ the cardinality of $\mathbf{c}(i)$, and $err(i,j)$ the percentage of scenes from class i wrongly classified as class j from Table 3.3. Each cell(i,j) representing a unique pair-wise emotion in the upper diagonal of Table 3.4 is then obtained using the formula

$$\text{cell}(i,j) = \frac{|\mathbf{c}(i)|\,err(i,j) + |\mathbf{c}(j)|\,err(j,i)}{|\mathbf{c}(i)| + |\mathbf{c}(j)|} \tag{3.15}$$

and then normalized by dividing it by the sum of all the cells in Table 3.4 and expressing it in terms of percentage such that all the percentages add up to 100%.

Of the twenty-one possible pairs of emotions, Sad-TA, Sad-Surprise, Fear-Surprise, Anger-Joyous, Anger-Surprise, Sad-Neutral and Anger-Fear are in descending order the seven pairs most culpable for errors. The confusion arises due to the frequent co-existence of these emotion pairs in cinema (i.e. the emotions of scenes that are labeled to contain two emotions usually belong to one of these seven emotion pairs). In some examples below, we illustrate typical instances of scenes where misclassification is common. For example, tenderness is often portrayed when someone comforts a despondent loved one. This implies that both TA and Sad emotion elements co-exist, thus increasing the classification error of the Sad-TA emotion pair if the classifier were forced to decide for only one of the emotions for such scenes. Large errors occur for the Sad-Surprise pair because many Surprise scenes are based on suspense, which occupies a very similar region in VA space (the low-arousal and low-valence region) as Sad scenes.

As for the third pair Fear-Surprise, these emotions are so intertwined in cinema that it is sometimes hard to even manually differentiate between them. Regarding the fourth pair Anger-Joyous, the difficulty arises because Joyous comedic scenes are usually slapstick in nature, and contain a fair amount of action elements inside, hence the confusion with Anger. In Chapter 2.5.1, the emotion pairs (Anger-Surprise) and (Anger-Fear) are two of the emotion pairs that have explicitly been identified by the cognitive perspective to overlap in VA space. Thus observation that these two emotion pairs are amongst the top seven pairs responsible for classification errors corroborates with emotion theory.

Finally, neutral scenes are dominated by short scenes with dialogue in dull or subdued tones, which are not uncommon in Sad scenes, thus complicating the

discrimination of the Sad-Neutral pair. Having discussed the difficult cases, it is nevertheless well to note that the confusion matrix indicates that most scenes are classified correctly. Despite the challenge of classifying every scene of the entire movies, and the significantly larger number and increased subtlety of emotional categories compared to existing works, the overall correct classification rate is 74.69%, or 85.82% if 'alternate selected' scenes are included (last row of Table 3.5). The second column under 'Alternate Selected' refers to those cases of dual label scenes whose dominant and alternate labels have received the second highest and highest probabilities respectively. Given the intrinsic ambiguity of these scenes and such a stringent criterion imposed, we believe that these so-called "alternate selected" scenes have been adequately classified and fully deserve a separate result category. Empirically speaking, the promising results suggest the classification of the affective categories is well-posed and separable using low level cues.

TABLE 3.3
Confusion Matrix for Extended Framework (%)

	Anger	Sad	Fear	Joyous	Surprise	TA	Neutral
Anger	69.38	3.93	5.62	8.15	7.30	0.84	4.78
Sad	2.66	61.13	3.65	3.99	10.96	11.96	5.65
Fear	7.57	1.08	83.78	0.54	6.49	0.00	0.54
Joyous	6.59	1.55	0.388	80.62	1.94	2.71	6.20
Surprise	6.16	11.23	9.42	1.81	65.22	3.26	2.90
TA	0.00	17.35	0.00	3.20	3.20	71.69	4.57
Neutral	4.50	7.19	1.57	5.84	5.40	4.05	71.46

TABLE 3.4
Confusion Matrix for Pairwise Affective Classification (%)

	Anger	Sad	Fear	Joyous	Surprise	TA	Neutral
Anger	0	3.76	6.61	7.57	6.56	0.52	4.44
Sad	0	0	3.19	3.19	11.68	15.81	6.75
Fear	0	0	0	0.47	7.97	0	1.22
Joyous	0	0	0	0	1.75	2.85	5.61
Surprise	0	0	0	0	0	2.97	4.02
TA	0	0	0	0	0	0	3.91
Neutral	0	0	0	0	0	0	0

The results also shed light on the relative influence of the audio and visual cues on classification. Rows 2 and 3 of Table 3.5 present the classification results using the audio and visual cues individually and jointly. Firstly, it is evident that combining both audio and visual cues together for classification significantly outperforms either of the cues individually. The results also corroborate our view that audio cues are far more informative than visual cues with respect to affective content, confirming our initial expectations. To the extent that the visual cues presented in this paper have captured the visual reality, we observe that there is a general lack of strong correlation between

the simple low-level visual cues presented here and the affective content. For instance, except at the extreme regions of the *HSL* color space, color does not correlate well with the affective content [19]. This lack of correlation is compounded by the fact that the director is constrained by the plot and general settings in the amount of freedom to set up visual environments that will evoke the desired moods.

However these difficulties do not seem to arise as severely for low level audio cues, where sound in film seems to be more immediately purposeful than visual details. In fact, the correlation between the low-level audio cues and the scene affect is in general so strong that unless there is a good reason for them to contradict (e.g. for comedic effect), the scene itself can easily be misinterpreted or appear jarring. Table

TABLE 3.5
Overall Classification Rate (%)

	Correct	Alternate Selected	Incorrect
Visual	42.86	9.87	49.27
Audio	61.39	10.34	28.27
Audio/Visual	74.69	11.13	14.18

TABLE 3.6
Ranking of Affective Cues

Cues	%	Cues	%
Silence Proportion	38.1	Environ Proportion	23.7
Joyous SAVE	32.1	Visual Excitement	20.2
Fear SAVE	28.7	Speech Proportion	20.2
Surprise SAVE	27.5	Median Lighting	19.7
TA SAVE	27.4	Shadow Proportion	19.4
Anger SAVE	27.4	Average Shot Length	17.6
Sad SAVE	26.2	Color Energy	17.4
Neutral SAVE	26.1	Music Proportion	16.7

3.6 ranks each cue according to the average rate of correct SVM classification between every pair of categories using only that cue. This table corroborates the finding that at the low-level, audio is more informative than visual cues. In particular, if one views absence of sound as an audio cue (a negative kind of audio cue), the top eight cues are all audio cues, with silence proportion as the most effective one. The sheer presence of absolute silence can be most dramatic and unsettling at times.

The performance of our algorithm compares favorably with the 78.7% reported by Kang [10], the only work we are aware that has performed affective classification on Hollywood scenes. However with regards to Kang's results, there are several important caveats. The test and training sets contain only selected scenes that are unambiguously manually labeled as one of only three classes: happy, sad or fear. The scenes were also selected from only six movie segments each lasting half an hour, as opposed to all the scenes in a movie.

3.5.3 Application

Machine understanding of the affective aspect of Hollywood multimedia can enhance and complement existing classification systems at several levels of resolution. Here we demonstrate applications at two levels: the more generalized movie genre level, and the more refined movie affective vector level (Figure 3.10). Other possible applications include using scene-level affective results for story unit extraction.

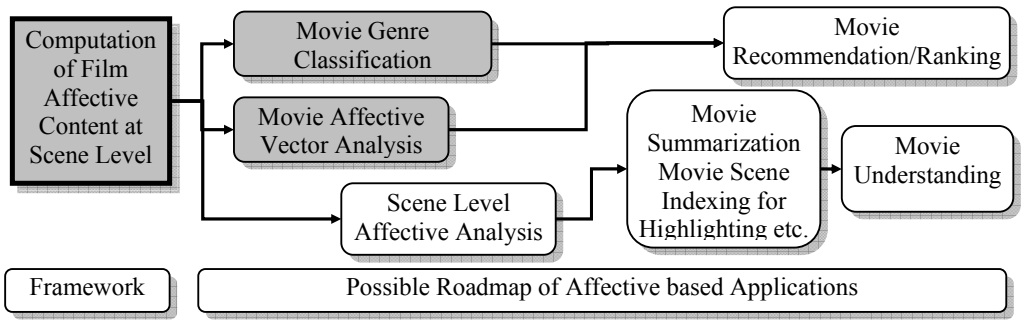


Figure 3.10 Illustration of possible roadmap for applications based on affective understanding in film. Shaded areas denote completed tasks.

3.5.3.1 Movie Genre Level

As it is, movie genres are sometimes too general to reflect the true character of a movie. For instance, genre labels seldom differentiate between comedic action and film noir action movies, or between tender drama and melodrama movies, although these differences substantially impact the movie experience. An obvious application of our work is to offer a more refined classification of any given movie and to detect dual genre movies, thus complementing existing genre classifications. In general, the genre of a movie can be largely determined by the proportion of time occupied by each of the affects. For example, a movie that has a significant amount of Fear and Surprise scenes is likely to belong to the horror genre. Therefore we let every movie be characterized by a movie affective vector (MAV), or $\mathbf{V}_i$,

$$\mathbf{V}_i = \frac{\left(\sum_{s=1}^{N_i} \mathbf{\Lambda}_{s,i} \tau_{s,i}\right)}{\left(\sum_{s=1}^{N_i} \tau_{s,i}\right)} \tag{3.16}$$

where i, s and N_i are the movie index, scene index and total number of scenes respectively. $\mathbf{\Lambda}_{s,i}$ is the vector of probabilities of each affective category for a scene

and $\tau_{s,i}$ is the duration of the sth scene in the ith movie. Effectively, V_i captures the affective content of a movie by weighing the affective vectors of the individual scenes comprising the movie with their duration. These scene affective vectors were already obtained by sending the scene feature vectors into the probabilistic classifier described in Chapter 3.4.

The notion of affective vector then can be readily extended to the genre, V_g, where the summation and normalization is carried out over each genre rather than a movie. The genre of a movie can then be determined by the distance measured between an MAV and that of a genre. We adopted the symmetrical Kullback-Leibler distance measure. Let the individual entries of the query and genre affective vectors be $V_{i,m}$ and $V_{g,m}$, indexed by m which runs over every affective category, and i, g which are the movie and genre indices respectively. The measure $M_a(i,g)$ is then defined as

$$M_a(i,g) = \sum_{m=1}^{N_i} \left(V_{i,m} \log\left(\frac{V_{i,m}}{V_{g,m}} \right) + V_{g,m} \log\left(\frac{V_{g,m}}{V_{i,m}} \right) \right) \tag{3.17}$$

A movie i is then assigned the genre g that returns the lowest $M_a(i,g)$, because lower $M_a(i,g)$ values indicate that similarity between a movie and the movies that make up the genre is strong. For the training, we take one movie out and train using the rest of the movies. The resultant classifier is then used to test the movie that has been taken out. This training-testing procedure is repeated for all the movies and 80.6% of all the movies are assigned the correct genre. All the "wrongly assigned" movies are listed in Table 3.7, where the 3rd and the 4th columns list the genres of these movies returning the lowest and second lowest $M_a(i,g)$. As can be seen, the genres returning the second lowest $M_a(i,g)$ values of all these movies correspond to the manual labels. This shows

TABLE 3.7
Movie Genre Classification Based on Scenes

Movie Title	Manual	1st Label	2nd Label
Jumanji	A	H	A
Lord of the Rings I	A	H	A
Saving Private Ryan	A	D/M	A
Up Close and Personal	R/C	D/M	R/C
Notting Hill	R/C	D/M	R/C
Life is Beautiful	D/M	R/C	D/M
Artificial Intelligence	D/M	A	D/M

The labels refer to the genres: Action(A), Horror(H), Romance/Comedy(R/C) and Drama/Melodrama(D/M).

that even for "wrongly assigned" movies, the manual labels of these movies have nevertheless received strong similarity values during testing.

Further inspection reveals that these results indeed reflect the dual nature of the majority of the movies in Table 3.7. For example, *Life is Beautiful* is actually a romance/comedy in the first half and a melodrama in the second half, while *Saving Private Ryan* is accurately described as a melodrama with a strong action element. On the other hand, *Artificial Intelligence* is correctly classified as an adventure (action) movie with much melodrama. Interestingly, although *Lord of the Rings I* is billed as an action movie, it has far more than its expected share of scary scenes. Classified by the algorithm as being most similar to the horror genre, the movie would warrant a caution for young children viewing it.

3.5.3.2 Movie Affective Vector (MAV) Level

At the next finer level of analysis, the MAV offers a more detailed picture of a movie. Due to the probabilistic inference framework adopted, the relative amounts of each affective component within a movie can now be estimated, as shown in Table 3.8.

This facilitates ranking of the movies according to a very useful and hitherto unimplemented aspect: its affective content. For instance, the affective vector of a movie can rank just how "happy" or "aggressive" etc. a movie is.

We survey Table 3.8 for broad quantifiable trends by genre, noting that the existence of characteristic genre MAV patterns underlies the consistency of the MAV. Aggression and Surprise feature prominently for the action genre while the D/M genre is dominated by TA and Sad elements. The R/C movies are marked by strong Joyous/TA affects while horror movies tend towards Fear/Surprise inducing scenes.

At the very top of the Aggression ranking list are *Speed* and *Starship Troopers*; the former is unrelentingly fast paced while the latter is extremely violent. Unsurprisingly, most other movies from the action genre followed closely, with the exceptions being *Saving Private Ryan*, which has strong melodramatic elements, and *Jumanji*, which being of the type "family entertainment", abstained from overt violence. Popular horror cinema can generally be differentiated by the main directing technique employed to induce fear: creating overtly threatening situations (Fear) or the more subtle tension (Surprise), which shows up clearly in their MAVs. *Ring* and *Gothika* are outright frightening while *The Others* depends far more on Surprise; the rest of horror movies possess a rather even mix of both elements.

As a rule of thumb, the summation of Fear and Surprise is a good indicator of the "scariness" of a movie. According to this indicator, *Ring* and *What Lies Beneath* would correctly be the scariest and mildest movies respectively. The same pattern appears for R/C movies, which generally belong to two groups: comedy/slapstick (Joyous) or sentimental (Tender Affection). Similarly, MAV can be used to classify these two groups and to rank them according to the summation of Joy and TA:

Bedazzled is slapstick, *My Best Friend's Wedding* is sentimental while *Notting Hill* aptly has the MAV of a subtle drama depicting a tortuous romance.

Besides using the MAV on its own for analysis, it can also complement the manually assigned genre of a movie if available, which provides the context for a more refined interpretation of the MAV. For example, because D/M movies tend to feature many dialogue scenes, it is understandable for such movies to obtain a high Neutral score. However if horror movies scored highly in the Neutral category, then such movies are not likely to be successful horror titles. Another example: having a high Surprise score for a romance/comedy genre movie usually indicates the presence of pleasant surprises while the same high Surprise score for a horror movie should be interpreted to imply the presence of unexpected shock and suspense.

From the aforementioned paragraphs, MAV analysis is able to yield broadly accurate ranking results according to different affects and even differentiate between different sub-genres, leading to automatic movie recommendation according to personalized affective preferences. To our knowledge, this is a capability not yet available in existing systems. With further investigation, more interesting and subtle patterns from MAV analysis are likely to emerge.

TABLE 3.8
Movie Level Affective Vector

Action	Agr	Sad	Fear	Joy	Sur	TA	Neu
The Fifth Element	31	10	1	9	32	5	12
Speed	69	8	8	3	4	3	6
Lord of the Rings I	19	11	25	9	28	4	4
James Bond (Golden Eye)	45	11	3	3	18	6	15
True Lies	41	8	2	7	14	5	23
Men In Black	33	3	1	2	42	1	17
Saving Private Ryan	24	26	3	5	10	19	13
Starship Troopers	60	7	2	10	4	3	13
Star Wars I	36	9	4	18	7	3	24
Waterworld	36	8	1	10	24	9	12
Jumanji	13	26	13	3	29	7	9
Drama/Melodrama (D/M)							
Forrest Gump	17	19	2	10	4	21	28
Magnolia	21	23	4	8	8	15	21
Ghost	23	12	9	3	6	22	24
Life is beautiful	21	17	2	37	6	9	7
City of Angels	11	33	3	5	8	26	15
Artificial Intelligence	18	32	8	2	17	9	15
The Sixth Sense	4	41	4	3	18	14	17
Horror							
Ghostship	9	8	34	3	20	5	22
Queen of the Damned	11	16	29	1	33	3	7
The Haunting	4	6	23	0	39	3	24
What Lies Beneath	4	30	14	2	22	10	17
The Others	9	26	11	2	43	6	3
Dream Catcher	19	7	42	6	6	3	17
Ring	2	7	77	1	4	3	6
Gothika	5	19	55	1	9	5	6
Legend of the Mummy	8	6	19	1	28	2	37
Romance/Comedy (R/C)							
There's Something About Mary	12	3	1	65	2	4	12
My Best Friend's Wedding	8	5	0	28	2	52	4
Up Close and Personal	13	14	2	7	5	38	21
Bedazzled	12	2	2	68	4	4	8
50 First Dates	7	5	2	61	2	17	6
Maid in Manhattan	8	5	0	22	1	47	16
Love Actually	3	8	0	32	2	46	9
Bruce Almighty	12	7	2	44	2	18	16
Notting Hill	12	16	1	16	5	21	29

The abbreviations for the emotions are respectively: Aggression (Agr), Sad (Sad), Fear (Fear), Surprise (Sur), Tender Affections (TA) and Neutral (Neu).

3.6 Conclusion

A complementary approach has been proposed to study and develop techniques for understanding the affective content of general Hollywood movies. We laid down a set of relevant and theoretically sound emotional categories and employed a number of low level features from cinematographic and psychological considerations to estimate these emotions. We discussed some of the important issues involved in automated affective understanding of film. We demonstrated the viability of the emotion categories and audiovisual features by carrying out experiments on large numbers of movies. In particular, we introduced an effective probabilistic audio inference scheme and showed the importance of audio information. Finally, we demonstrated some interesting applications with the resultant affective capabilities.

Much work remains to be done in this largely unexplored field. Firstly, with regards the shortcomings of our work, the small proportion of scenes that are wrongly classified shows up the inherent limitation of low-level cues (especially visual) in bridging the affective gap. Therefore in the immediate future, we intend to implement more complex intermediate-level cues to further improve present results. Secondly, the existence of multiple emotions in scenes requires a more refined treatment. Finally, we will also investigate the possibility of finer sub-partitioning of the present affective categories, as well as further scene affective vector level analysis.

CHAPTER IV

MOTION BASED OBJECT SEGMENTATION

4.1 Introduction

In contrast to other types of media, the defining aspect of film, tellingly known also as moving pictures, is the presentation of information through the use of visual motion. Due to its saliency in describing certain interesting semantic concepts from largely directed video genres like film, sports and program shows, motion has gradually gained recognition as a potent feature for video indexing. This is well attested by the inclusion of motion descriptors in the MPEG-7 [103] standard like motion activity, camera movement, trajectory, and parametric motion, for the purposes of similarity-based video retrieval [80], video abstraction [121] and structuring video data [122]. A small sample of further example applications include identification of certain types of baseball plays [96], football plays [75] and presentation annotation [97], underlining the immense potential of motion for film shot semantic indexing.

As the narrative video genre with one of the most sophisticated and mature traditions, film is embedded with different forms of semantics at various levels of abstraction. However it contains less apparent structure than news, talk shows and sports, thus attempts to extract film semantics using superficial motion-based low level computable features are met with limited success. In particular, these efforts are hampered by two issues: insufficient incorporation of domain specific knowledge to

suggest appropriate and more meaningful film shot categories, and lack of intermediate level computable visual features to facilitate such classification. In cinematography, a pivotal set of informal production rules we term *directing grammar*, further elaborated in the next chapter, governs the relationship between a subset of film shot semantics and camera related attributes, especially motion.

Our work addresses the first issue by explicating the intimate multifaceted relationship that exists between film shot semantics and computable visual features by mining this directing grammar. For instance, detection of tracking operations in a shot likely indicates the presence of subject(s) of interest, which is of strong indexing value. Another common directing rule relies on the camera distance from the subject(s) of interest to re-orientate the audience or adjust the relative emphasis between the subject and surrounding environment, indirectly revealing some shot composition information and directing intentions as indexing cues. Exploiting this grammar enables us to propose a sufficiently high level semantic taxonomy for film shots to be of interest to users.

On the second issue, we develop novel independent motion detection capabilities, specially adapted for cinema considerations, enabling us to formulate and compute effective motion-based descriptors that map to high level film shot semantics. This enables a compact number of salient and robust directing descriptors to accurately capture the directing semantics inherent in a film shot. Existing applications that can benefit from these systems include automated film analysis [69], editing [70], film structure creation [71], indexing [72] and video abstraction/summarization [121] for both specialized commercial and mass-consumer applications.

In general, every shot contains at least one subject or place of interest, which we call the Focus-of-Attention (FOA), and one main directing intention, conveyed mainly by manipulating the audience viewpoint in three different ways: 1) camera motion operations, 2) framing choice in the portrayal of FOA and finally 3) the duration that FOA(s) remain on-screen. This facilitates the central goal behind the viewpoint manipulation: namely to direct visual attention [121], a process whereby the director concentrates the interest of the viewer to closely observe an object or place. It is this act of attention, which privileges certain intended viewer experience and interpretation of what is observed, that ultimately defines the directing semantics of a given shot. Recovering the relevant computable directing descriptors to decipher these semantics necessitates accurate motion segmentation and foreground-background identification capabilities.

Thus in this chapter, we will detail and present the results of a novel motion segmentation technique, specially designed for film shot semantics recovery, which forms the algorithmic core of the motion based Hollywood film shot indexing framework. The development and demonstration of this framework, grounded in cinematographic domain knowledge, will be discussed in the next chapter.

4.2 Motion Segmentation Literature Review

Motion segmentation is commonly viewed as the process of mapping sets of pixels (or supports) to different motions, and can be delineated into two major approaches. The first approach, also referred to as the top-down approach, seeks to recover the dominant motion, or the motion to which a majority of the pixels conforms. Pixels that conform to the dominant motion are grouped under one motion label while

other are treated as outliers. This process continues iteratively on the remaining outliers until most, if not all pixels, belong to a computed motion. The work by Odobez [48], which exemplifies this approach, returns excellent performance for the scenario with a dominant background and one motion-wise coherent foreground object. However there is no in-built competitive process between different motion hypotheses, and motion configurations deviating from this scenario can cause difficulties.

The second approach seeks to estimate all motions and their respective supports simultaneously and in turn has three major variants. The first variant, also known as the bottom-up approach, estimates a large number of motions, one for each small image area, before merging patches that are similar in motion [49][52][53][54][55] (typically done using k-means clustering). One of the earliest and most representative works of this approach belongs to Wang and Adelson [49], who merged image patches exhibiting similar affine motions. This approach is attractive both in term of computation and flexibility, since redundant computation is avoided and the number of motions can vary as circumstances demand. However the criterion of merging must be carefully formulated to prevent wrong assignments. A unique example of this variant proposed by Shi and Malik [50][51] used a global graph partitioning method known as the Normalized Cut to best partition pixels into two dissimilar sets, which can be viewed as a merge that finally results in two sets.

The second variant is the level set formulation where the boundaries of curves evolve according to certain partial differential equations [56][57][58] to group together areas of motion homogeneity, replacing active contours as the descriptors of motion boundaries. It does not require much parameterization and can adapt easily to changing

topologies. However the final solution does depend on initialization and areas labeled under the same motion must be contiguous.

The last variant is the general Expectation Maximization (EM) framework, a popular technique [60][61][62][63][64][65][66]. In recent years, a large number of such works represent pixels or regions within the image with graphs, in order to find the motion labels of these image parts within the general EM framework. The Markov Random Field (MRF) [131] has emerged as a leading approach to model the motion labels of the image parts and their interactions amongst themselves and the observations. A critical strength of EM is the presence of competition amongst different motion hypotheses for image parts that encourages label assignments most capable of explaining the observations. An interesting innovation of EM by Sawhney and Ayer even prunes excessive motion models using a size-of-model criterion under a Minimum Description Length framework [62]. However there are some drawbacks to EM. The number of hypothesized motions must be fixed *a priori*, which is not realistic under many circumstances. Also, computation increases linearly with the numbers of hypothesized motions over the entire image, even if some motions are localized to only a small area of the image.

Besides analyzing algorithms according to the relationships between the motion estimation process and support areas, a variety of constraints are being utilized in the cost function. The most universally used constraint, also known as the data term, is the likelihood of the resultant displaced intensity differences (DID) assuming correct motion estimation; a constraint directly related to the assumption that intensity is preserved across frames. The second term, or spatial continuity term, encodes the belief that regions from the same object should be spatially connected, and encourages

neighboring regions to share similar labels [55][65]. Another popular term is the temporal continuity term, which seeks to preserve the same labels in regions across time; Mezaris *et. al.* [54] is notable in this regard for considering the temporal track for the entire shot, as opposed to neighboring frames, before making the decision on semantic object level segmentation for each frame. A few works have also explicitly incorporated an appearance term [50] that depends on texture etc. Finally, the latest works have begun to consider depth ordering, via the observation of occlusion, as a constraint on motion segmentation. Some have used this as a separate post-processing step [49], however, Tweed and Calway [55], Drummond [60] and Torr et al [64] have integrated this constraint into the segmentation process itself.

4.3 Our Motion Approach

Recovering shot semantics in the extraordinarily diverse and dynamic film domain environment is a most challenging matter. We eschew purely statistical approaches that do not possess explicit object concepts, and are therefore severely handicapped by the inability to truly identify foreground and background. We also avoid a solely trajectory based approach, which depends heavily on continuous tracking, and are far too fragile for a movie environment where objects can vary tremendously in size, number, speed and even undergo occlusion. Instead, we have elected to incorporate key observations from directing grammar to guide our approach. The resultant incorporation of directing grammar leads to a novel motion segmentation algorithm specially adapted for film shot semantics recovery.

In the film domain, it is observed that the dominant frame motion has almost invariably belonged to either an unobtrusive background, or the FOA. Exploiting this

key fact, our approach ultimately labels all image parts with only two motion labels: the dominant motion and “all-other-motions”. Then, it sets upon the critical task of identifying the FOA by deciding whether the FOA should belong to the dominant or “all-other-motion” label. To accomplish this, we assume a) the dominant motion belongs either to the FOA or background and 2) the FOA is always in the foreground. These reasonable assumptions, elaborated in the next sections, allow us to recover the dominant motion without being concerned with the variable and possibly large number of independent motions during the motion segmentation process.

Our contribution here lies in the incorporation of directing grammar into the design of a motion segmentation scheme, based on using MRF to model the motions that different parts of each frame take on. This scheme features novel integrated occlusion reasoning to tell apart the FOA from non-FOA areas, precluding the need for separate pre-processing and ad-hoc occlusion detection methods [49], while allowing the foreground and background to be correctly identified at the global level without making the somewhat broad and common assumption that the dominant motion is always background [65]. Furthermore, intrinsic to the segmentation is the ability to track the evolution of viewer attention at the shot level by approximating how humans direct their attention towards FOA, allowing the creation of more informative descriptors.

To ensure robustness and relevance in the dynamic film environment, we utilize a dominant motion framework where key parameters adapt automatically to the dominant motion speed for optimal segmentation. Unlike most motion algorithms, it relies exclusively on the most informative pixels - the edge pixels as opposed to all pixels - to extract motion information. The details for each step of the motion

segmentation algorithm are provided in the rest of the chapter. A system overview of our motion segmentation algorithm is illustrated below (Figure 4.1).

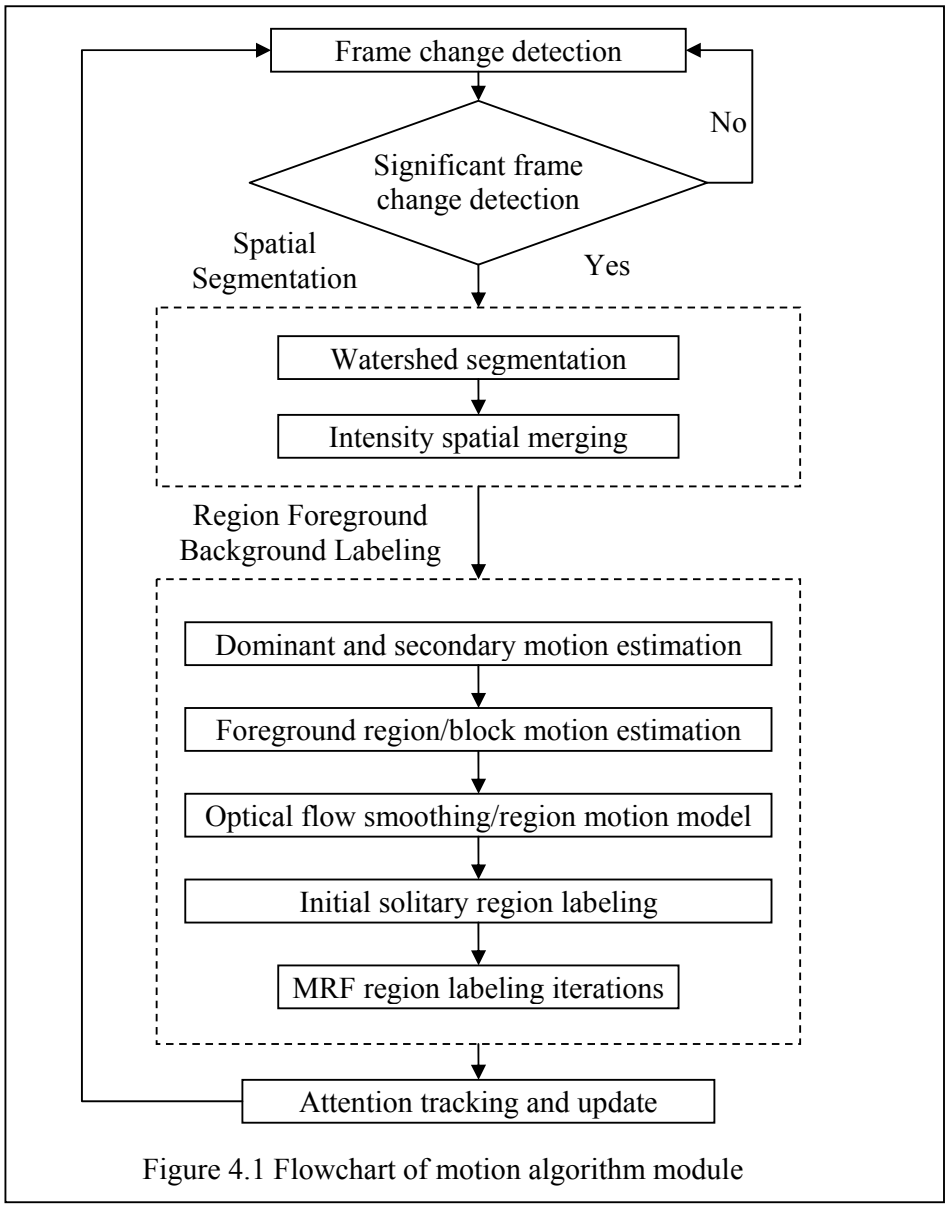


Figure 4.1 Flowchart of motion algorithm module

4.4 Region Segmentation and Merging

Since a prerequisite to motion segmentation is the presence of motion itself, we have implemented a frame change detection step that computes the intensity differences between two frames, and skips to the next frame if the proportion of pixels exhibiting an absolute difference larger than 6 is below 5%. This improves computation time, and more importantly segmentation accuracy, for shots with long periods of stationarity.

To facilitate motion segmentation, we utilize a region based representation of the image frame by grouping spatially connected pixels by intensity into regions. This is because motion estimation performed over multiple pixels can utilize their collective information at the region level, enabling higher robustness to image noise, aperture problem and occlusion compared to the pixel level. Another important advantage is the computational savings that accrues from dealing with a small number of regions as opposed to the substantially larger number of pixels. Here we provide a brief description of our region segmentation implementation using the watershed segmentation algorithm [67][68].

A suitable one channel intensity function F is first chosen to represent the frame. The gradient image G, a 2D matrix comprising the Sobel gradient magnitudes of every single pixel in F, is then computed. Being the derivative of F, G can be treated as a topographical surface that shows up the locations of sharp intensity differences in F, which in turn constitute the boundaries of the segmented regions. From this perspective, watershed regions are actually areas of nearly uniform intensity values in F bounded by sharp intensity changes. In order to partition F, we employ the

waterfall simulation method on G, which assigns every single pixel in F to a watershed region based on where its steepest descent path in G leads to. The watershed segmentation thus partitions F into the set of watershed regions denoted by $\mathbf{R}_{ws} = \{R(1), R(2), \ldots, R_{ws}(N_{ws})\}$, where N_{ws} is the original number of watershed regions. A drowning step is finally carried out to automatically merge adjacent pixels separated by a boundary weaker than a certain pre-determined threshold Th_{drown}.

Due to the critical dependency of spatial segmentation on chromatic intensities, we evaluated several permutations of color spaces. The luminance channels Y and L are respectively chosen from the *YCbCr* and *CIE Lab* color spaces as the intensity function F for the segmentation. Additionally, to incorporate segmentation information from other color channels, the maximal value of all three color channels for each pixel is also used for the intensity function F. This brings the total number of candidate intensity functions to four: Y, L, *maxLab* and *maxYCbCr*.

Subject to the condition that object boundaries are not violated, we judge the color space that 1) yields fewer regions during segmentation and 2) requires less computational resources as the most optimal. From experiments performed to obtain the optimal segmentation parameters and F (Figure 4.2), it is clear that Y and L produce significantly more regions that *maxLab* and *maxYCbCr*. This can be attributed to the fact that not all object boundaries are captured in the luminance component alone. Due to the linearity of the *maxYCbCr*, which bestows large speed advantages over *maxLab*, we have adopted *maxYCbCr* for watershed segmentation, using Th_{drown}=10.

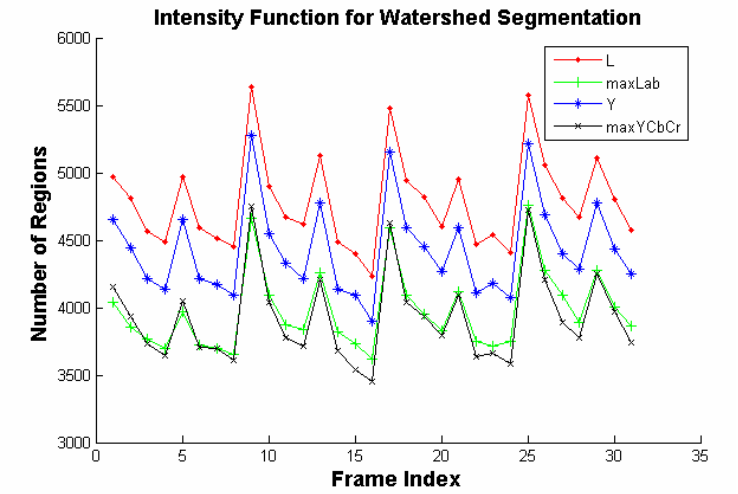


Figure 4.2 Segmentation regions for different color-spaces. In the first 31 frames of *Foreman* sequence, the strong chromatic similarities between the foreman's yellow helmet and the wall directly behind it allows us to adjust parameters associated with the four intensity functions till the helmet is correctly segmented.

The resultant over-segmented image preserves the object boundaries at the cost of significant computation cost arising from having to deal with numerous small regions. More importantly, the larger perimeter-to-area ratio of small regions increases the likelihood of higher motion compensated difference (MCD) errors and inaccuracies in the subsequent motion estimation processes. Hence there is a need to reduce this original set of regions through region merging.

Thus we define two inter-region measures based on chromaticity and edge strength to aid the spatial merging process. Let the color centroid of any region with index *i* be denoted by the vector $\{I_i^1, I_i^2, I_i^3\}$, which represents the three color channels of the color space used during region merging. The color centroid differences between two regions *i* and *j* are then defined as

$$I_{Diff}(i,j) = \sqrt{(I_i^1 - I_j^1)^2 + (I_i^2 - I_j^2)^2 + (I_i^3 - I_j^3)^2} \tag{4.1}$$

To also quantify the edge strength between two regions, we define an edge strength difference measure

$$G_{Diff}(i,j) = \frac{\sum_{k=1}^{Nij} G(bdr_k(i,j))}{N_{ij}} \tag{4.2}$$

where $bdr_k(i,j)$ is the set of pixels along the borders of regions i and j, k indexes into this set and N_{ij} is the cardinality of the set. With these two measures, the algorithm for region merging is as follows:

1) Sort the set of regions by size and arrange them into a linked list for fast indexing.

2) Starting from the smallest region R_i, merge it with the neighboring region with which it has the lowest $I_{Diff}(i,j)$, subject to:

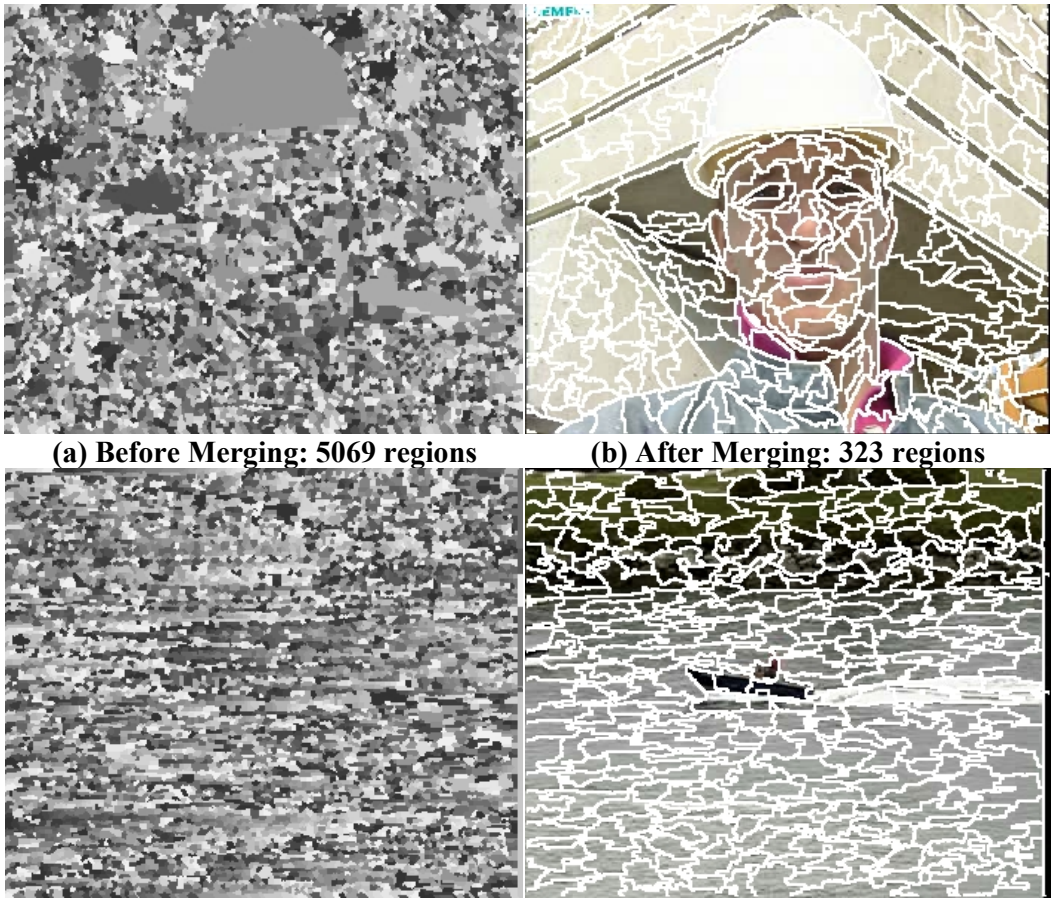
$[(R_i(size) < Th_{Size\_small})] \vee$

$[(Th_{Size\_small} < R_i(size) < Th_{Size\_med}) \wedge (I_{Diff}(i,j) < Th_{I\_Diff\_med}) \wedge (G_{Diff}(i,j) < Th_{G\_Diff\_med})] \vee$

$[R_i(size) > Th_{Size\_large}) \wedge (I_{Diff}(i,j) < Th_{I\_Diff\_large}) \wedge (G_{Diff}(i,j) < Th_{G\_Diff\_large})]$

3) This process of merging continues until the smallest region is above Th_{Size_large}, or until no further merging can occur.

We denote the final set of regions that remain after merging as $\mathbf{R} = \{R_1, R_2, \ldots, R_N\}$, where N is the total number of final regions. From systematic experiments, we found *YCbCr* color space to be the optimal color space for region merging. The optimal values for region merging empirically set to $Th_{I\_Diff\_med}$=20, $Th_{G\_Diff\_large}$=5, $Th_{I\_Diff\_small}$=10 and $Th_{G\_Diff\_small}$=40. Th_{Size_small}, Th_{Size_med} and Th_{Size_large} are set at (1/2000), (1/500) and (1/15) of the size of the frame. This results in an average reduction by a factor of 18 in the number of regions (Figure 4.3).



(a) Before Merging: 5069 regions **(b) After Merging: 323 regions**

(c) Before Merging: 7106 regions **(c) After Merging: 345 regions**

Figure 4.3 Region merging. A drastic reduction in regions for the watershed segmentation after the merging procedure for frames 1 of the *Foreman* (a,b) and *Coastal* (c,d) sequence.

4.5 Core Motion Estimation Algorithm

The core motion estimation algorithm used for all motion estimation purposes thereafter is introduced in this section. To accomplish the essential task of motion estimation in image sequences, a parametric motion model estimation algorithm, incorporated with robust estimator (maximum likelihood) capabilities in a multi-resolution framework, is used. The attractiveness of using a parametric model lies with its ability to compute a dense and good approximation of the optical flow utilizing only a small number of parameters (maximum 6 for affine flow) within a region, reasonably

assumed to be under coherent motion. In addition to simplifying computation by reducing the number of unknowns, it suppresses minor specific motions in favor of a more general and useful motion description. The use of a robust estimator (Tukey biweight estimator) further minimizes the contribution of outliers in occlusion cases to return more reliable optical flow estimates. Finally the multi-resolution scheme enables large motions to be computationally feasible in a gradient-based motion estimation framework. The following sub-sections describe the algorithmic details [47] involved in motion estimation.

4.5.1 Parametric Motion Model

Motion is represented with the class of 2D polynomial motion models of the point coordinates (x, y) in the image plane up to the affine model, which is the first order polynomials in x and y. Using matrix notation, a model **M** is expressed as

$$\mathbf{V}(X_i) = \begin{bmatrix} u(X_i) \\ v(X_i) \end{bmatrix} = \mathbf{D}(X_i)\mathbf{M} \quad \text{and} \quad \mathbf{D}(X_i) = \begin{bmatrix} 1 & x_i & y_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_i & y_i \end{bmatrix} \tag{4.3}$$

where it is linear with respect to the n motion parameters $\mathbf{M}^t = (a_1, a_2, \ldots, a_n)$, and $X_i=(x_i, y_i)$ refers to the spatial image position of a point while $\mathbf{V}(X_i)$ refers to the flow vector at X_i. The affine flow motion model (n=6) is the motion model we utilize because it exhibits a good tradeoff between complexity and ability to represent motion sufficiently accurately (e.g. translation, rotation, scaling, and deformation). It describes the flow field w.r.t. the unknowns $\mathbf{M}^t=[a_1,a_2,a_3,a_4,a_5,a_6]$ as

$$\begin{aligned} u(X_i) &= a_1 + a_2 x_i + a_3 y_i \\ v(X_i) &= a_4 + a_5 x_i + a_6 y_i . \end{aligned} \tag{4.4}$$

To confer onto the optical flow estimation some adaptability to global illumination changes, the Brightness Constancy Equation (BCE) can be written as

$$r_i = I(X_i + \delta X_i, t + \delta t) - I(X_i, t) + \varsigma \delta t \tag{4.5}$$

where r_i is the residue of X_i and ς denotes the global change in illumination. By noting that $\delta X_i = \mathbf{V}(X_i)$ and taking $\delta t = 1$ to simplify the notation, r_i can be re-expressed as

$$r_i = (I(X_i + \mathbf{D}(X_i)\mathbf{M}, t+1) - I(X_i, t) + \varsigma) \tag{4.6}$$

4.5.2 Multi-Resolution Least Square Estimation

An estimate of the motion field is computed by obtaining the set of optimal motion parameters for the motion parametric model that minimizes the displaced intensity differences (DID), which is the residue, between two consecutive frames. A multi-resolution least square estimation scheme, which uses an incremental estimation of the motion model through course-to-fine refinement, is thus used to minimize the following sum of squared difference errors, or error function:

$$E(\mathbf{\Phi}) = \sum_{X_i \in S} (r_i)^2 = \sum_{X_i \in S} (I(X_i + \mathbf{D}(X_i)\mathbf{M}, t+1) - I(X_i, t) + \varsigma)^2 \tag{4.7}$$

where S denotes the region of support the algorithm uses to perform motion estimation. Let $\mathbf{\Phi} = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ \varsigma]$, $\hat{\mathbf{\Phi}}, \hat{\mathbf{M}}, \hat{\varsigma}$ be the current estimates of $\mathbf{\Phi}, \mathbf{M}, \varsigma$ and $\Delta\mathbf{\Phi}, \Delta\mathbf{M}, \Delta\varsigma$ the incremental change to be computed, then

$$\begin{gathered} \mathbf{\Phi} = \hat{\mathbf{\Phi}} + \Delta\mathbf{\Phi} \\ \begin{bmatrix} \mathbf{M} \\ \varsigma \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{M}} \\ \hat{\varsigma} \end{bmatrix} + \begin{bmatrix} \Delta\mathbf{M} \\ \Delta\varsigma \end{bmatrix} \end{gathered} \tag{4.8}$$

It is well known that with the increment so defined, Taylor's expansion can be used to expand the residue to the first order of $\mathbf{I}$ around point $X_i + \mathbf{D}(X_i)\hat{\mathbf{M}}$ at time $t+1$ to give

$$
\begin{aligned}
r_i' &= \mathbf{A}_i \Delta\mathbf{\Phi} - B_i \\
A_i &= \left[\nabla I(X_i + \mathbf{D}(X_i)\hat{\mathbf{M}}, t+1)\mathbf{D}(X_i), \quad 1 \right] \\
B_i &= I(X_i, t) - I(X_i + \mathbf{D}(X_i)\hat{\mathbf{M}}, t+1) - \hat{\varsigma}.
\end{aligned}
\tag{4.9}
$$

Hence by casting the error function as a quadratic error measure minimization

$$
E(\mathbf{\Phi}) = \sum_{X_i \in S} (r_i')^2 = \sum_{X_i \in S} (\mathbf{A}_i \Delta\mathbf{\Phi} - B_i)^2 \tag{4.10}
$$

we can obtain the incremental solution via the iterative least square approach:

$$
\Delta\hat{\mathbf{\Phi}} = \left[\sum_{X_i \in S} (A_i)^t A_i \right]^{-1} \sum_{X_i \in S} (A_i)^t B_i. \tag{4.11}
$$

This incremental estimation is embedded in a coarse-to-fine refinement scheme where several levels of Gaussian low-pass pyramids of each image are used. At the coarsest level, no estimation is available. However for a typical image size of 352x288, having four image pyramid levels typically ensures that the displacements are small enough for the BCE to approximately hold true, thus a null motion initialization is used. For each level, successive iterations using equations (4.8)-(4.11) are performed until the incremental estimate is too small or the limit for number of iterations is reached. Then, the estimated parameters are transmitted to the finer level, where the refinement process starts again. These cycles of iterations are repeated until the finest level is reached.

4.5.3 Robust Outlier Estimation

Outliers caused by noise, occlusion and violation of motion model assumptions are ameliorated using the Tukey-biweight robust estimator ρ to minimize the contribution of outliers. The shape of robust estimator ρ with derivative ψ

$$\psi(r_i', C) = \begin{cases} r_i'(1-(r_i'/C)^2)^2 & if \ \ |r_i'| < C \\ 0 \end{cases} \tag{4.12}$$

allows the error penalty, which is the value returned by the robust estimator ρ, to be capped at a constant for residue r_i above a certain value C=8, thus mitigating the undue effect of outliers on the parameter computation. Incorporating this robust estimator into the error function gives:

$$E(\mathbf{\Phi}) = \sum_{X_i \in S} \rho(r_i') = \frac{1}{2} \sum_{X_i \in S} w_i (r_i')^2 \quad and \quad w_i = \frac{\psi(r_i')}{r_i'}. \tag{4.13}$$

where w_i represents the measure of membership of each pixel to the current motion parameters **M**. The weight w_i, whose range is between [0,1], can be directly interpreted as the likelihood of pixel i being assigned the correct optical flow. The support S is over the set of edge pixels lining the perimeter of some regions returned by the watershed segmentation. The regions included in the support S depend on which particular stage of the hierarchical optical flow estimation process they are in (see next section). Expressing the error function in terms of motion parameters gives

$$E(\mathbf{\Phi}) = \frac{1}{2} \sum_{X_i \in S} w_i (r_i')^2 = \frac{1}{2} \sum_{X_i \in S} w_i (\mathbf{A}_i \Delta\mathbf{\Phi} - B_i)^2 \tag{4.14}$$

while the incremental update equation becomes

$$\hat{\Delta\mathbf{\Phi}} = \left[\sum_{X_i \in S} w_i (A_i')^t A_i' \right]^{-1} \sum_{X_i \in S} w_i (A_i')^t B_i'. \tag{4.15}$$

In practice, within every iteration loop of the incremental motion estimation, there exists another loop that computes the w_i and r_i consecutively and iteratively using equations (4.9), (4.12)-(4.15) till both values stabilize.

4.6 Hierarchical Optimal Flow Estimation

Motion segmentation and attention tracking require accurate and robust computation of optical flow between two frames, especially in the presence of occlusion. Due to the tremendous computation needed to recover dense optical flow, we have adopted a hybrid approach combining top down and bottom up techniques. Basically, this separates the entire motion estimation procedure into various stages where motion is estimated in the following order: frame, large region and finally the block level. Estimation is carried out at higher levels first, and stepped down to lower levels only for pixels that do not yet conform to any of the previously computed motions. Correspondingly, the motion models used for motion estimation also steps down in complexity with the decrease in support pixel area in order to achieve significant time savings and motion estimation robustness. This information is eventually smoothed and accumulated to give a robust inference of the optical flow for each region. The entire process of motion estimation is illustrated in Figure 4.5.

Different from the vast majority of existing motion algorithms ([60] is one of the few exceptions), we have chosen to utilize only the pixels lining the perimeter of the regions, which we call edge pixels, for this task. A key difference between us and [60] is our usage of the DID of these edge pixels as opposed to the Euclidean distances between them for motion estimation. The exclusive use of edge pixels at the global level is motivated by three reasons. Firstly, due to their high intensity gradients, edge pixels contain much more reliable motion information compared to the smooth homogenous pixels comprising the region interior, which cannot constrain the interpretation of motion. Secondly, using only edge pixels typically consume only two-

thirds of the computation time compared to using all the pixels. Thirdly, and most importantly, with a sufficient number of edge pixels, using edge pixels exclusively generally gives more accurate motion segmentation compared to using all pixels, especially in the presence of occlusion. Because of their high intensity gradients, edge pixels require highly accurate motion estimation to minimize DID, and are less likely to fall into a spurious local minimum.

4.6.1 Region/Block motion estimation

Initially, motion estimation is performed over all edge pixels to obtain the dominant motion. Let the average weight of the edge pixels (defined in Equation 4.13) of region i under the current motion estimate be $w(i)$, then all regions with $w(i)>0.7$, which implies a high likelihood that the pixels conform to the current motion, are assigned to the current motion model and their pixels are retired from subsequent motion estimation process. This first recovered motion is designated as the dominant motion and the process is iterated repeatedly for other secondary major motions until the region areas under these secondary motions fall below 15% of the image area.

For the remaining regions that are not yet assigned a motion model due to their smaller support, it is more robust to adopt a more localized motion estimation approach. The motion model used has also been stepped down in simplicity to one that estimated only translation and divergence [47]. We adopt this treatment for every of the unassigned regions whose areas exceed the size of three 32x32 blocks.

By this stage, the only regions with no motion models are those that are neither large nor conformed to the dominant and major secondary motions. We proceed to assign a fitting motion model to each of these regions with the following procedure.

First, the frame is tessellated into 32x32 blocks and motion estimation is performed in each individual block if at least 10% of the block is occupied by unassigned edge pixels. We use only the edge pixels of unassigned regions within that block, and the estimated motion is then assigned to the entire block. Given the modest support and local nature of each block, we step down to a uniform 2D translation model because it is adequate and provides improved stability over more complex motion models.

Due to the occurrences of occlusion or multiple motions within a block, the motion computed from the block estimation step may be spurious or wrongly assigned. Thus each block is further tessellated into 8x8 "block-lets" to further refine motion assignment. The idea is to assign to the block-let an optimal motion model from the motions of the 32x32 blocks neighboring the block-let. Since optical flow is usually smooth, we can assume the true motion model of any block-let *i* is within the candidate set **CS**(*i*) of motion models obtained from the nine nearest 32x32 blocks (see Stage 3 of Figure 4.5). All the unassigned pixels – not only edge pixels – of each block-let *i* are greedily initialized to the motion from **CS**(*i*) that maximizes the average pixel weight. Let **NBLM**(*i*) be the set of optimal motion models currently adopted by the eight neighboring block-lets of block-let *i* and $w_{blk}(a,m_b)$ the average pixel weight of block-let *a* under motion m_b. The optimal motion selection process is then carried out for each block-let in raster-line order to select the motion *j* from **CS**(*i*) that best minimizes the following objective function *OBJ(i)* over all *j*:

$$\begin{aligned} OBJ(i) &= \varsigma(w_{blk}(i,m_i) - w_{blk}(i,m_j)) + \sum_{k}^{\mathbf{NBLM}(i)} f_{mtnDist}(m_j, m_k). \\ f_{mtnDist}(m_j, m_k) &= [1\ 1\ bs\ bs\ bs\ bs] \left| m_j - m_k \right| \end{aligned} \tag{4.16}$$

where ς=30, *bs* is the block size=32, *j* indexes into **CS**(*i*), current motion $m_i = [a_{1,i}\ a_{2,i}\ a_{3,i}\ a_{4,i}\ a_{5,i}\ a_{6,i}]^T$ and candidate motion $m_j = [a_{1,j}\ a_{2,j}\ a_{3,j}\ a_{4,j}\ a_{5,j}\ a_{6,j}]^T$. The first term of the objective function is a data term to encourage selecting a motion for the block-let that maximizes its average pixel weight (Stage 4 of Figure 4.5), while the second term $f_{mtnDist}(m_j,m_k)$ is a smoothness term promoting smoothness of neighboring motion models by penalizing dissimilarity between neighboring block-let motion models. This entire raster-line optimal motion model selection process is iterated seven times in all, and the block-let motions of the last iteration are accepted as the final motions.

Finally, to select the best motion model for those regions which have been tessellated in blocks, we define the notion of a region motion consistency *(RMC)* measure, which computes the consistency of any particular motion model in a region (Stage 5 of Figure 4.5). The idea is to implement a voting scheme where the motion most consistent with other models within the same region is chosen as the optimal region motion model. Let a region *i* contain N_m different motion models in the set **RM**(*i*) amongst its pixels, and *q* indexes **RM**(*i*). Let the membership of pixels belonging to the motion m_q have a cardinality of $|m_q|$. Then the *RMC(i,p)* for motion model *p* in region *i* is defined as:

$$RMC(i,p) = \sum_{q \in \mathbf{RM}(i)}^{N_m} \min(|m_p|, |m_q|) * reg_{mtnDist}(m_p, m_q)$$
$$reg_{mtnDiff}(m_p, m_q) = \begin{cases} 0 & f_{mtnDiff}(m_p, m_q) \geq 2 \\ 1 - (f_{mtnDiff}(m_p, m_q)/2) & f_{mtnDiff}(m_p, m_q) < 2 \end{cases} \quad (4.17)$$

The motion model p within **RM**(i) with the highest *RMC(i,p)* is thus adopted as the region motion model. With this, robust and accurate dense optical flow is obtained (Figure 4.4).

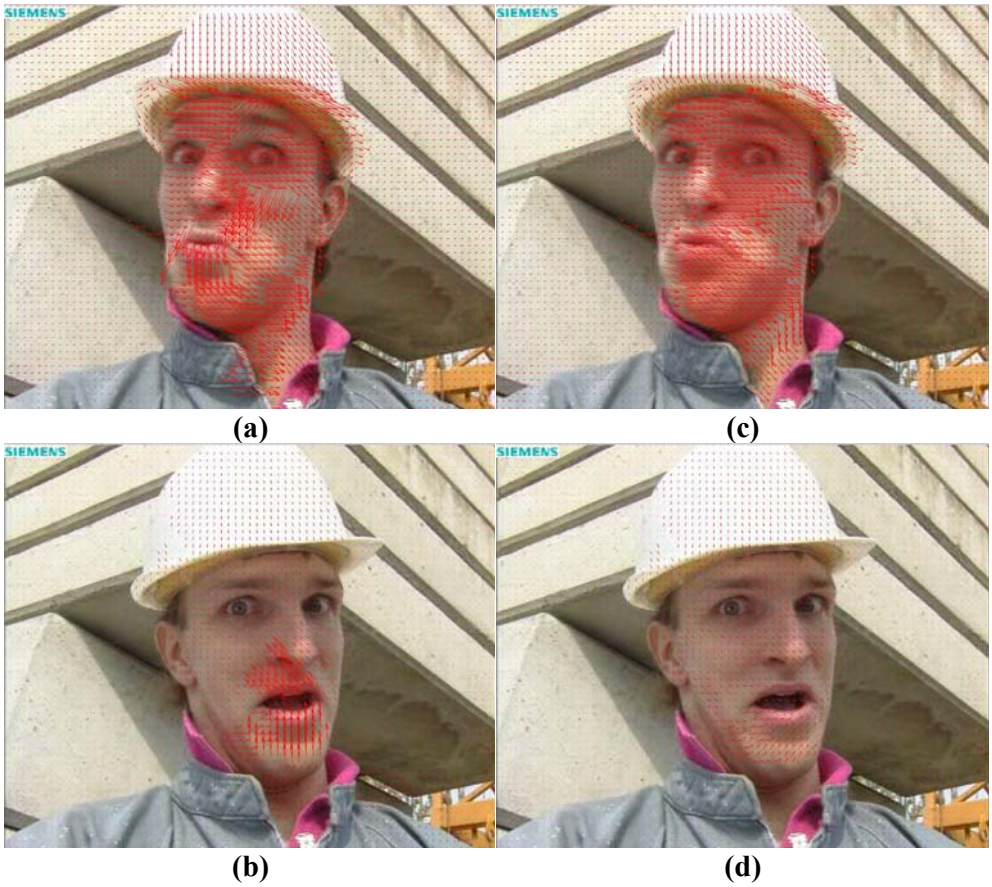
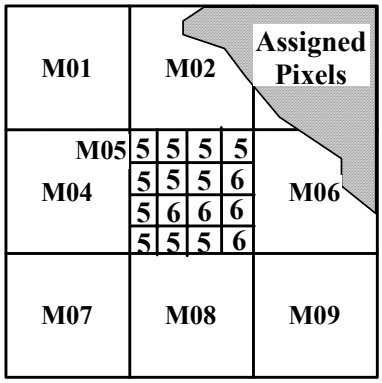


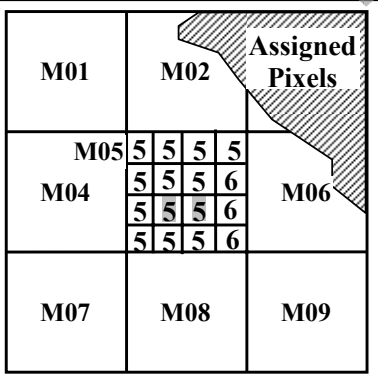
Figure 4.4 Optical flow smoothing. Optical flow results with smoothing (c,d) and without smoothing (a,b), where (c,d) exhibit smoother optical flow conforming far more closely to the real motion.

Stage 1: Initial dominant (black) and non-dominant area segmentation.

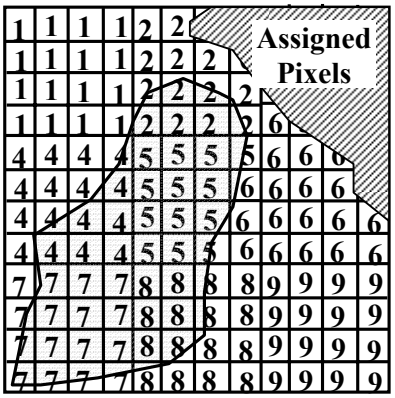
Stage 2: Those regions not tentatively classified under the dominant, major secondary or large region motions are tessellated into 32x32 blocks for parametric motion estimation. Only edge pixels are used for motion estimation. However the computed motion models are assigned to all unassigned pixels within the block. Different motion models are indicated with different colors.



Stage 3: The 32x32 block M05 is divided into 16 8x8 block-lets. Motion models that maximize block-let weights are selected from amongst motions of neighboring blocks.



Stage 4: Iterative smoothing is performed at the 8x8 block-let level. Shaded block-lets show changed motion models after smoothing.



Stage 5: For each region (e.g. the shaded shape), region model selection is performed by selecting the motion model of the block-lets within the region that fits best with other motion models in the region.

Figure 4.5 Illustration of the optical flow computation process.

4.7 Motion Segmentation with Markov Random Field (MRF)

Known for their ability to capture spatial relationships in 2D image processing, the MRF is a graphical model in use for several decades in various applications ranging from image restoration [117] to motion segmentation [65]. Due to the Hammersley-Clifford theorem [119], which proved the equivalence between Gibbs random fields and MRF, MRF has been established as a numerically tractable and attractive tool to model a large variety of non-causal processes.

The MRF offers a formal approach capable of easily incorporating a priori domain knowledge, capturing spatial relationships and modeling a variety of dynamics. In our case, it allows us to incorporate a few crucial cinematographic constraints previously alluded to. Firstly, we assume the Focus of Attention (FOA) to be in the foreground. While there are some exceptions, it holds true for the vast majority of cases. Secondly, it is assumed that the dominant motion is sufficiently accurate to describe either the FOA or the background. Although the background usually conforms quite rigidly to one motion, this may not be the case for the FOA, which, for instance, may be a walking human with multiple articulated hand motions. For our shot indexing purpose, it is nevertheless sufficient to identify the dominant motion exhibited by the human body as foreground, while leaving out the swinging hands.

In accordance with these two constraints, we cast the region labeling problem as one where each region is labeled either as foreground or background. We also need to determine if the region's motion belongs to that of the "dominant motion" or "all other motions". Thus at the global level, the region labeling process has to test two hypotheses, *Hypo1* (dominant=*BG*, others=*FG*) and *Hypo2* (dominant=*FG*,

others=*BG*). The hypothesis with the higher probability is taken to be the correct interpretation.

Applying MRF modeling [61][65][131] to our region labeling problem, we represent the N regions $\mathbf{R}=\{R_1, R_2, R_3,\ldots, R_N\}$ as the set of MRF sites, which are defined on a popular neighborhood system Ξ where physically adjacent sites are neighbors. A hypothesis of this MRF thus comprises of the set of random variables, or configuration $\xi=\{\xi_1, \xi_2,\ldots, \xi_N\}$, where ξ_i can take on either the *FG* or *BG* label for region i, as well as the variable H, which can take on *Hypo1* or *Hypo2*. ξ is said to be a Markov random field on $\mathbf{R}$ with respect to a neighborhood system Ξ if and only if two conditions are satisfied: 1) positivity: $P(\xi=\xi')>0,\ \xi'\in$ space of all possible ξ and 2) markovianity: $P(\xi_i \mid \xi_{\{\mathbf{R}\}-i}) = P(\xi_i \mid \xi_{\Xi_i})$ where $\{\mathbf{R}\}$-i is the set difference, $\xi_{\{\mathbf{R}\}\text{-}i}$ denotes the labels of sites $\{\mathbf{R}\}$-i, while $\xi_{\Xi i}$ denotes the labels of sites neighboring site i on neighborhood system Ξ**.** Positivity is a very mild condition usually assumed in practice, while the markovianity assumption is satisfied by modeling likelihoods that depend only upon neighboring sites, as elaborated in the next sub-section.

Let the observation $\mathbf{O}=\{O_1, O_2, O_3, \ldots, O_N\}$ be the set of individual features O_i observed for region R_i. The solution we seek is the optimal configuration ξ and hypothesis H that maximizes the MAP (maximum *a posteriori*) $P(\xi|\mathbf{O},H)$. On the assumption of a uniform prior $P(\xi)$, and given a constant evidence $P(\mathbf{O},H)$, MAP is proportional to the likelihood $P(\mathbf{O},H|\xi)$ and is expressed in the Bayesian framework as

$$P(\xi \mid \mathbf{O}, H) = \frac{P(\mathbf{O}, H \mid \xi)P(\xi)}{P(\mathbf{O}, H)} \propto P(\mathbf{O}, H \mid \xi) \tag{4.18}$$

In conjunction with the Hammersley-Clifford theorem [119], the MAP can be expressed as a Gibbs distribution

$$P(\boldsymbol{\xi} \mid \mathbf{O}, H) \propto \frac{e^{-U(\mathbf{O},H|\boldsymbol{\xi})}}{Z_H} \tag{4.19}$$

where Z_H is the partition function and is constant for a specific H, while the MAP is maximized by minimizing the likelihood energy function $U(\mathbf{O},H|\boldsymbol{\xi})$. $U(\mathbf{O},H|\boldsymbol{\xi})$ comprises the intra-region and inter-region interactions captured by the local clique potentials V^D, V^S and V^A to reflect our data, spatial and attention constraints. Let the set of all possible singleton cliques be C1 and pair-wise cliques be C2 in this neighborhood system, then

$$\begin{aligned} U(\mathbf{O}, H \mid \boldsymbol{\xi}) &= E_{data} + E_{spatial} + E_{attention} \\ &= \kappa_d \sum_{\{i\} \in C1} V_1^D(\xi_i, \mathbf{O}, H) + \kappa_a \sum_{\{i\} \in C1} V_1^A(\xi_i, \mathbf{O}, H) + \kappa_s \sum_{\{i,j\} \in C2} V_2^S(\xi_i, \xi_j, \mathbf{O}, H) \\ &= \sum_i^N \left(\kappa_d V_1^D(\xi_i, \mathbf{O}, H) + \kappa_a V_1^A(\xi_i, \mathbf{O}, H) + \kappa_s \sum_{j \in Neighbor(i)} V_2^S(\xi_i, \xi_j, \mathbf{O}, H) \right) \end{aligned} \tag{4.20}$$

where κ_d, κ_s and κ_a are constants that control the relative importance of these three sets of energy potentials. The clique potentials V_1^D and V_1^A are defined on singleton cliques, while V_2^S is defined on pair-wise cliques, in accordance to the neighborhood system. Higher order cliques are not explored due to the exponential increase in complexity, and more importantly, because the interactions we want to model (as elaborated in the next sections) are fully met with cliques defined up to the pair-wise level.

4.7.1 Data Term

The MRF solves two problems simultaneously: which regions conform to the dominant motion (the motion recovered first in the preceding section), and whether these regions belong to the foreground or background. For the first task, it is readily

determined by considering for each edge pixel how well the inter-frame image intensity is preserved under this dominant motion. We recomputed the weight w of all the region edge pixels using this dominant motion, and obtain the average edge pixel weight $w(i)$ of region i. This recomputed $w(i)$ is utilized as the main observation to determine the data energy term. In theory, $w(i)$ follows distinct likelihood distribution curves for both dominant motion regions (high likelihood) and non-dominant motion regions (low likelihood), and the point where they intersect would be the optimal place to read off a decision threshold, which can be used for data energy term computation.

However it is known empirically that this decision threshold varies strongly with two factors: global speed and the intensity gradient magnitude of each region. High global speed causes blurring of region boundaries, causing spuriously low $w(i)$. At the same time, low intensity gradient magnitude for region i, $av_{Grad}(i)$, leads to inaccurately high $w(i)$. Hence these factors have to be flexibly compensated for to obtain robust region labeling.

We approach this problem by computing an adaptive decision threshold $EQ_w(i)$ for each region i based on these two factors. First, for a frame with average dominant speed $avGS$, we define various speed levels at spd_{lo}=2, spd_{mid}=4, spd_{hi}=12, and various intensity gradient magnitude levels at $grad_{lo}$=20 and $grad_{hi}$=50. Then a global adjustment GS_{adjust} that is adaptive to the various speed and gradient magnitude levels can be computed with the following equation:

$$GS_{adjust} = \begin{cases} 0.1 & avGS \le spd_{lo} \\ 0.1 * \dfrac{(avGS - spd_{lo})}{(spd_{mid} - spd_{lo})} & spd_{lo} < avGS \le spd_{mid} \\ -0.1 * \dfrac{(avGS - spd_{mid})}{(spd_{hi} - spd_{mid})} & spd_{mid} < avGS \le spd_{hi} \\ -0.1 & avGS \ge spd_{hi} \end{cases} \tag{4.21}$$

The global adjustment is then added to $EQ_w(i)$ in a step that also compensates for possibly low *avGrad(i)*:

$$EQ_w(i) = \begin{cases} EQ_{wc} + GS_{adjust} & avGrad(i) > grad_{hi} \\ EQ_{wc} + GS_{adjust} + (avGrad(i) - grad_{lo})/200 & grad_{lo} < avGrad(i) \le grad_{hi} \\ EQ_{wc} + GS_{adjust} + 0.15 & avGrad(i) \le grad_{lo} \end{cases} \quad (4.22)$$

where EQ_{wc}=0.7 is the equilibrium weight constant. After determining the suitable decision threshold $EQ_w(i)$ for every region *i*, the data energy potential is calculated as:

$$V_1^D(\xi_i, \mathbf{O}, H) = \begin{cases} -\frac{1}{2} \cdot \frac{w(i)}{EQ_w(i)} & w(i) \le EQ_w(i) \\ -\left(\frac{1}{2} + \frac{1}{2} \cdot \frac{w(i) - EQ_w(i)}{1\text{-}EQ_w(i)} \right) & w(i) > EQ_w(i) \end{cases} \quad (4.23)$$

where κ_d=5. Note that the $V_1^D(\xi_i,\mathbf{O},H)$ is designed such that both foreground and background are equi-probable when $w(i)==EQ_w(i)$. In this context, $EQ_w(i)$ plays a very important role in the computation of an almost continuous and variable data energy function, as opposed to a binary function.

The heart of the occlusion handling mechanism revolves around the computation of *w(i)*, where edge pixels are selectively summed depending on the hypothesized relative depth order of any region *i* with its adjacent regions. Edge pixels tend to be occluded whenever occlusion occurs, and as Bergen [115] observed, given the correct motions for any two regions, error density is often high on the occluded side of an edge, and low on the occluding side. For instance, edge pixels belonging to *FG* labeled regions are always taken into account when computing the data energy because these pixels are supposed to be unoccluded, and hence observable. However the edge pixels belonging to a region labeled as *BG* and bordering regions labeled as

FG are excluded from computation because they have a high likelihood of being occluded, as per the hypothesis.

By varying the behavior of how edge pixels are used to compute the data energy between adjacent regions with differently labeled depths, it enables the modeling of occlusion into the MRF process (Figure 4.6) instead of an ad hoc or complicated multi-motion occlusion/depth order handling process. Let $bdr_w(i,j)$ denote the total weights of the set of edge pixels along the borders of regions i and j, N_i the number of regions bordering i and let $av()$ be the average operator, then

$$\begin{aligned} w_{fg}(i) &= av(\sum_{j}^{N_i} bdr_w(i,j)) \qquad \xi_i = FG \\ w_{bg}(i) &= av(\sum_{j}^{N_i} bdr_w(i,j) \cdot f_{occ}(i,j)) \quad \xi_i = BG \\ f_{occ}(i,j) &= \begin{cases} 0 & \xi_i = BG, \xi_j = FG \\ 1 & \text{other cases} \end{cases} \end{aligned} \tag{4.24}$$

Finally, the value of $w(i)$ is calculated differently depending on both the value H, and the label ξ_i of region i. Note that $w(i)$ is computed for both hypotheses *Hypo1* and *Hypo2*.

$$w(i) = \begin{cases} w_{bg}(i) & H = Hypo1, \xi_i = BG \\ 1 - w_{bg}(i) & H = Hypo1, \xi_i = FG \\ 1 - w_{fg}(i) & H = Hypo2, \xi_i = BG \\ w_{fg}(i) & H = Hypo2, \xi_i = FG \end{cases} \tag{4.25}$$

Figure 4.6 Comparison of occlusion energy. Segmentation results with occlusion (c,d) and without occlusion (a,b) factored into the energy calculations. Background regions in (a,b) near the helmet and neck that encounter occlusion are wrongly classified as foreground.

4.7.2 Spatial Term

The spatial energy potential, consisting of pair-wise cliques, expresses the a priori assumption that regions belonging to the same label (*FG* or *BG*) have similar colors, and tend to cluster together. Thus this energy encourages the formation of compact and adjoining foreground and background boundaries. Let N_{ij} be the perimeter length and $I_{Diff}(i,j)$ the L_2 distance between the color centroids in *YCbCr* space of regions *i* and *j* respectively. Then

$$V_2^S(\xi_i,\xi_j,\mathbf{O},H)=\begin{cases}-f_{se}(i,j)N_{ij} & \xi_i=\xi_j\\ 0 & \xi_i\neq\xi_j\end{cases} \tag{4.26}$$

$$f_{se}(i,j)=\begin{cases} clr_{lo} & I_{Diff}(i,j) < t_{c,lo} \\ clr_{lo}+\dfrac{clr_{hi}-clr_{lo}}{t_{c,hi}-t_{c,lo}}(I_{Diff}(i,j)-t_{c,lo}) & t_{c,lo} \le I_{Diff}(i,j) \le t_{c,hi} \\ clr_{hi} & I_{Diff}(i,j) > t_{c,hi} \end{cases} \quad (4.27)$$

where κ_s=2 and $f_{se}(i,j)$ is a color centroid similarity measure to encourage adjacent regions with similar color to take on similar labels. Referring to the above equations, clr_{hi}=2 and clr_{lo}=0.5 control the range of the spatial energy potentials, while $t_{c,lo}$=20 and $t_{c,hi}$=80 control the thresholds that determine how well a given $I_{Diff}(i,j)$ value satisfies the color similarity assumption between adjacent regions *i* and *j*.

4.7.3 Attention Term

As discussed previously, the directing semantics of a shot is subtly influenced by the skilful direction of the viewer's attention. By correctly identifying the FOA in each frame, and tracking the number of times different areas receive attention throughout the shot, an "attention signature" can be composed from such information for each shot. This is motivated by the expectation that different classes of shot semantic have their own characteristic attention signatures.

To model this attention process, we use two 2D image buffer the size of the image frame: Rec_{att} to record the net duration a pixel has been classified as FG in the most recent 25 frames, subject to a ceiling of T_{att_span}=25, and $Hist_{att}$, to record the total number of times the pixel has been classified as *FG* in the shot. The value of T_{att_span} for Rec_{att} is equivalent to approximately one second in duration, considering the standard frame rate of 25 fps, and is chosen to model the persistence behavior of attention span remaining on an area that has stopped moving before it fades and all focus is transferred to other moving areas.

We denote $avR_{att}(i)$ and $avHist_{att}(i)$ as the average value of the pixels of region i in the current Rec_{att} and $Hist_{att}$ respectively. We exclude from computation the pixels in the current frame that are not mapped to by the optical flow from the previous frame pair. This is because such areas tend to be those that have just newly appeared and would spuriously pull down $avR_{att}(i)$ and $avHist_{att}(i)$ if they were included. By the same token, any region with unmapped area exceeding 50% of its own area are probably newly appearing background and will have its corresponding pixel locations in both buffers set to zero. All pixels in both buffers for each region are then updated to their respective $avR_{att}(i)$ and $avHist_{att}(i)$ values.

We posit that the longer any object has been moving in recent memory, the likelier it is to continue receiving attention. To model this phenomenon and encourage smoothness in labeling along the temporal dimension, a threshold T_{att}=5 is introduced which encourages *FG* labeling for a region with $avR_{att}(i)$ above T_{att}, while penalizing a region being classified as *FG* if its $avR_{att}(i)$ is below T_{att}. Note that T_{att} is set much lower than T_{att_span} to model the assumption that FOA attracts viewer's attention faster than it is relinquished. Furthermore, we compute the attention energy potential only after a burn-in period of the first T_{att_span} number of frame pairs; this ensures we have sufficient evidence from past frames to compute the attention energy potential. Expressing the above modeling assumptions, with κ_a=1.5, the attention energy potential term can be written as

$$V_1^A(\xi_i, \mathbf{O}, H) = \begin{cases} -(avR_{att}(i) - T_{att})/(T_{att\_span} - T_{att}) & \xi_i = FG \\ 0 & \xi_i = BG \\ 0 & if\ \#\ frames < T_{att\_span} \end{cases} \quad (4.28)$$

4.7.4 Optimal Hypothesis and Region Labels

The process of finding the optimal configuration $\boldsymbol{\xi}$ and hypothesis H that minimizes $U(\mathbf{O},H|\boldsymbol{\xi})$ is separated into two stages. In the first stage, the hypothesis H is fixed either as *Hypo1* or *Hypo2* to ensure computation stability for the second stage, where actual iterative energy minimization is carried out. The configuration $\boldsymbol{\xi}$ is initialized in a greedy manner such that the labels maximize the solitary clique (i.e. non-pair-wise) data and attention energy potentials. Then, at every iteration to minimize $U(\mathbf{O},H|\boldsymbol{\xi})$ by adjusting $\boldsymbol{\xi}$, the total energy potential for both labels $\{FG, BG\}$ for all regions are computed, and the region that decreases $U(\mathbf{O},H|\boldsymbol{\xi})$ the most with its alternative label will switch its current label. This iterative minimization process, also known as the Highest Confidence First (HCF) method [118], continues until $U(\mathbf{O},H|\boldsymbol{\xi})$ reaches a stable minimum or a certain number of iteration, in our case $7N$, has passed.

At the conclusion of the separate MAP maximization iterative processes for both *H=Hypo1* and *H=Hypo2*, the "Weighted-Likelihood" (WL) energy $U_{WL}(\mathbf{O},H|\boldsymbol{\xi})$ is computed from the different final configurations $\boldsymbol{\xi}$ obtained under both H labels. $U_{WL}(\mathbf{O},H|\boldsymbol{\xi})$ weighs the likelihood of each region i by its visual presence, as modeled by its number of edge pixels $\gamma(i)$. A WL partition function $Z_{WL,H}$ is computed under pseudo-likelihood assumptions [131], defined as the simple product of the conditional likelihood, and in the large lattice limit is shown to converge to the true likelihood.

$$U_{WL}(\mathbf{O},H\mid\boldsymbol{\xi})=\sum_{i}^{N}\gamma(i)\begin{pmatrix}\kappa_d V_1^D(\xi_i,\mathbf{O},H)+\kappa_a V_1^A(\xi_i,\mathbf{O},H)\\ +\kappa_s\sum_{j\in Neighbor(i)}V_2^S(\xi_i,\xi_j,\mathbf{O},H)\end{pmatrix}$$
$$Z_{WL,H}=\exp\left(-\sum_{i=1}^{N}\sum_{\xi_i=\{BG,FG\}}U_{WL}(\mathbf{O},H\mid\xi_i,\{\boldsymbol{\xi}-\xi_i\}\right) \tag{4.29}$$

The WL MAP $P_{WL}(\boldsymbol{\xi}|\mathbf{O},H)$ is finally computed for both hypotheses, and the configuration $\boldsymbol{\xi}$ and hypothesis H responsible for the higher $P_{WL}(\boldsymbol{\xi}|\mathbf{O},H)$ are taken to be the truth labels.

$$P_{WL}(\boldsymbol{\xi} \mid \mathbf{O}, H) = \frac{e^{-U_{WL}(\mathbf{O},H|\boldsymbol{\xi})}}{Z_{WL,H}} \tag{4.30}$$

With the final region labelings, the two attention buffers for the kth pixel of region i are updated respectively as:

$$\begin{aligned}
Hist_{att}[k] &= \begin{cases} Hist_{att}[k] - \delta_{att} & \xi[k] = BG \\ Hist_{att}[k] + \delta_{att} & \xi[k] = FG \end{cases} \\
\delta_{att} &= \begin{cases} 1 & \left|Hist_{att}(i) - Hist_{att}[k]\right| \le T_{att} \\ 1 + floor\left(\left|avHist_{att}(i) - Hist_{att}[k]\right| / T_{att}\right) & \left|Hist_{att}(i) - Hist_{att}[k]\right| > T_{att} \end{cases} \\
\mathrm{Re}c_{att}[k] &= \begin{cases} \mathrm{Re}c_{att}[k] - \delta_{att} & \xi[k] = BG \\ \mathrm{Re}c_{att}[k] + \delta_{att} & \xi[k] = FG \end{cases} \\
\delta_{att} &= \begin{cases} 1 & \left|\mathrm{Re}c_{att}(i) - \mathrm{Re}c_{att}[k]\right| \le T_{att} \\ 1 + floor\left(\left|avR_{att}(i) - \mathrm{Re}c_{att}[k]\right| / T_{att}\right) & \left|\mathrm{Re}c_{att}(i) - \mathrm{Re}c_{att}[k]\right| > T_{att} \\ 0 & \mathrm{Re}c_{att}[k] == 25 \end{cases}
\end{aligned} \tag{4.31}$$

Finally, the values in both buffers Rec_{att} and $Hist_{att}$ are shifted according to the motion model assigned to each pixel. Locations that are unmapped due to uncovering are indicated as such while locations with more than one value mapped to it will accept the higher value. Occasionally spurious motion estimation and segmentation introduce inconsistencies into both attention buffers. To ameliorate this, we perform a 1-neighbor memory diffusion process on both Rec_{att} and $Hist_{att}$, using the 3x3 weighing kernel of [1, 1, 1, 1, 8, 1, 1, 1, 1] over pixels that belong to the same region to smooth the buffers and to strengthen the maintenance process. Figure 4.7, 4.8 illustrate the results of using the occlusion handling mechanism to identify the foreground and background correctly.

(a) frame 1 **(b) frame 25**

(c) frame 75 **(d) frame 100**

Figure 4.7 Identifying foreground and background. In an extremely fast moving action sequence from "The Dreamcatcher", the camera zooms in onto an F16, changing its size drastically. Despite this, the global FG/BG segmentation module is able to successfully distinguish between the two, unlike algorithms that simply associates the dominant motion with the background.

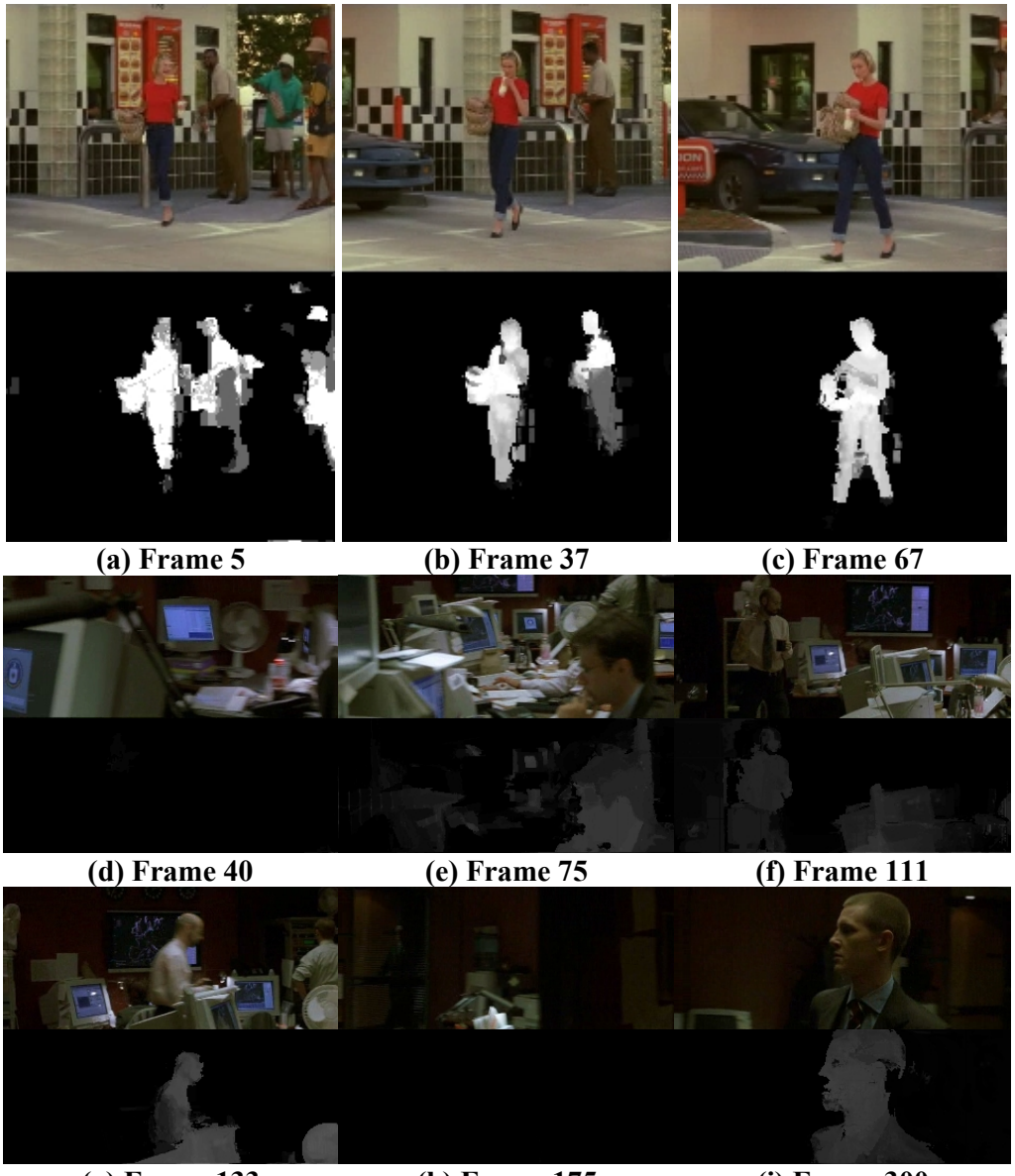
(a) Snapshot 1

(b) Snapshot 2

(c) Snapshot 3

(d) Snapshot 4

Figure 4.8 Snapshots taken from one of the famous scenes of "The Fellowship of the Ring". The characters, upon which the director directs our focus, are each larger than the background they have temporarily occluded.



(a) Frame 5 **(b) Frame 37** **(c) Frame 67**

(d) Frame 40 **(e) Frame 75** **(f) Frame 111**

(g) Frame 133 **(h) Frame 175** **(i) Frame 300**

Figure 4.9 Attention signature maps for two sequences (a-c) and (d-i). Whiter areas indicate higher attention intensity. Note the attention intensity rises and ebbs accurately according to the location of the FOA at the moment.

Figure 4.9 shows some examples of attention signatures. Figure 4.10 and Figure 4.11 illustrate the performance of the proposed motion segmentation algorithm with a large variety of realistic Hollywood shots, each with different motion content, characteristics and challenges. On a Pentium IV 3.4Ghz processor, the un-optimized

C++ algorithm takes an average of 1.26s to compute dense optical flow and foreground/background segmentation for a 352x288 frame. It is noted that for non-action normal paced shots, frame skipping is frequent and can decrease the total amount of expected processing time by half and even more.

Figure 4.10 Segmentation results from “There’s something about Mary”. Note the wide variety of camera distances present within these shots.

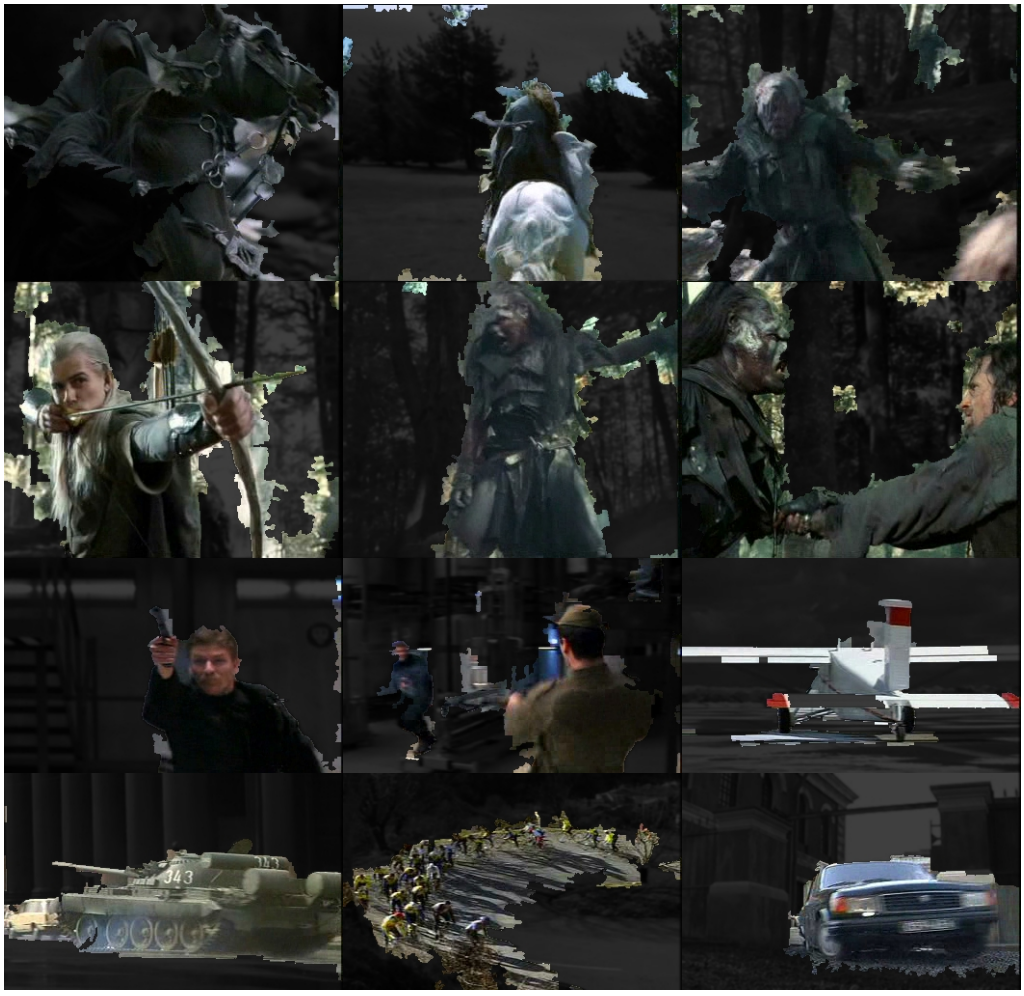


Figure 4.11 Segmentation results from the action movies “The Fellowship of the Ring” and “James Bond: Golden Eye”. Presented are the segmentation results of some of the most furious action sequences in either of these movies.

4.8 Difficulties Encountered by Motion Segmentation Module

Due to the extremely wide domain of Hollywood shots filmed under the most diverse circumstances possible, it is inevitable that motion segmentation will under some circumstances produce segmentation that does not conform to the human perception of the focus of attention.

One of the difficulties is associated with the confusion between foreground and background regions, typically for cases where the background is extremely bland and small in area compared to the foreground. Without effective optical constraints, the background tends to be assigned the wrong motion, causing errors in distinguishing between foreground and background. An example would be an extremely close up face shot framed by a bland wall by the sides.

The second class of difficult shots occurs whenever there is temporary occlusion of some tracked foreground object (i.e. a man who is tracked throughout the shot may be momentarily occluded with someone else walking in front of the camera). This wipes the accumulated memory away from the occluded foreground completely, rendering the memory values in the last frame unreliable as a guide of the on screen duration of an object.

The third class of difficult shots arises whenever the assumption of affine motion model for the background is violated, which is common for panoramic shots covering a wide range of depths, and even indoor shots. Another somewhat related manifestation of this problem is a miscellaneous and rare group of shots featuring highly non-rigid motion (swirling water, burning fire etc.) or special lighting effects.

Finally, some shots are so low in average intensity that even humans will find difficult to determine the exact optical flow. These shot usually appear in the horror genre, where it is not infrequent to encounter shots with totally dark background, or in effect, no background. For the purpose of this work, we have excluded shots that are too dim.

4.9 Conclusion

Discovering the directing intentions behind film shots potentially leads to a wealth of semantics necessary for various video content management and processing applications. In tackling this challenge, we have formulated a unique approach based on a pivotal observation from film grammar: namely the manipulation of the viewer's visual attention is what ultimately defines the directing semantics of a given shot.

To capture the salient information behind this attention manipulation process, we proposed a novel edge-based MRF motion segmentation technique, specially adapted for film shot semantics. This technique is capable of identifying the Focus-of-Attention (FOA) areas accurately by utilizing edge pixels to model occlusion explicitly. The elegant integrated occlusion reasoning dispenses with the need for ad-hoc occlusion detection, allowing the foreground and background to be correctly identified at the global level without making somewhat unrealistic assumptions on the background, as do other related works in the Hollywood domain.

The segmentation process inherently tracks FOA areas with attention maps and recovers accurate optical flow, which are vital to the eventual computation of effective and robust directing descriptors to extract shot directing semantics for indexing purposes. The motion segmentation is robust in the dynamic film environment, where key parameters adapt automatically to the shot characteristics for optimal segmentation. Furthermore, it does not make unwarranted nor restrictive assumptions on the size, number and speed of independently moving objects. Experiments show the algorithm performs satisfactorily on real life Hollywood shots, despite some difficult shot types, which we hope to solve in future works using non-motion a priori knowledge.

CHAPTER V

FILM SHOT SEMANTICS USING MOTION AND DIRECTING GRAMMAR

5.1 Introduction

Film directing grammar, used interchangeably with "directing" henceforth, is one of the most crucial set of production rules underlying the movie making process, due to its critical role of conveying director intentions through specific camera motions and viewpoint attributes. Though subtle, it exerts surprising amount of influence on how the viewer perceives and experiences the movie through three major ways.

Firstly, directing plays an instrumental role in focusing the attention in order to cue the viewer in on important details or objects. For instance, to strongly imply the presence of interesting FOA (Focus of Attention), a number of motion-related tasks such as tracking are routinely executed by the director. Secondly, directing prescribes different camera distances for framing viewpoints, which in turn provoke different subjective responses, an example being the well known rule that close-ups tend to have more emotion impact than long distance shots. Finally, it is able to change the perception of the passage of time; a prime example being the pace of change of camera viewpoint and motion. As the above examples illustrate, directing grammar implies a strong correspondence between a certain set of movie-related semantics and motion-related computable descriptors, which we term "directing descriptors". This correspondence can in turn be harnessed to map the directing descriptors, a hitherto

superficially utilized source of information, onto semantic level knowledge, as shown in Figure 5.1. Movie shots are passed through the motion segmentation algorithm for extraction of the relevant directing descriptors. These descriptors are in turn fed into the SVM classifier to infer what we term the shot "directing semantics".

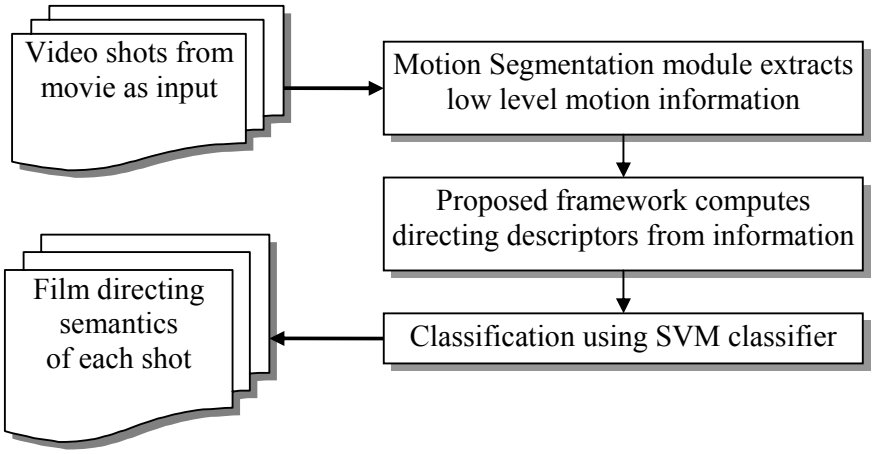


Figure 5.1 Flowchart of system overview.

Most prior motion based indexing works are limited to the sports domain [73][74][75][76][77], whose organized structure allows easier application of motion to recover semantics. In contrast, the use of directing grammar on a comprehensive basis to extract semantics in the more challenging film domain is seldom addressed, if at all directly, having been restricted to the extraction of a few simple types of camera motion such as pan, tilt and zoom. Others have concentrated on motion description capabilities that fall into the three major methods based on motion trajectory [78], motion activity [79] and statistical modeling [80][82][124][125]. However these works are more concerned with motion based retrieval, whose frameworks are generally not optimized with any domain in mind, as opposed to the semantically specialized motion based indexing, which explicitly models the target semantic categories. [81] and [83] have proposed using generic frameworks to recover high level semantics; however the

generality of the framework and low-level cues are not specifically tailored to mine semantic information embedded in film directing structure at the shot level.

In this chapter, we address the above issues by firstly proposing a qualitative film directing semantics taxonomy - organized using certain vital film directing elements - that articulates a set of the most significant vocabulary of directing semantics. Secondly, guided by directing film grammar, we formulate effective directing descriptors capable of recovering their corresponding semantics. Finally, we demonstrate our system with experimental results.

In determining the scope, parameters and domain of the work, we consider several issues. A common adage in film directing states that "every shot and cut fulfils a purpose". Since the shot is the natural level of abstraction for directing grammar based analysis to yield significant semantics, we seek to recover this purpose, or at least some meaningful characteristics at the shot level, for indexing purposes. Due to more elaborate setups, cinematographic motion as prescribed by directing tends to find fuller expression in Hollywood films; hence our focus on them. However we emphasize that directing grammar is also employed in the production of the vast majority of fictional video narratives ranging from dramas to mini-series, thus securing wide domain for the application of our work especially in automated film analysis [69], editing [70], film structure creation [71], indexing [72] and video abstraction/summarization [121]. To our knowledge, no work in the film domain recovers such a comprehensive set of film shot semantics, or even use directing grammar based methods exclusively. Though motion is the modality under investigation, our framework easily allows other modalities or even forms of semantics

(e.g. affective – Chapter 2) to complement and enhance its film shot semantics indexing capabilities.

5.2 Literature Review

Vision based document query systems can be organized according to whether they are primarily designed for indexing or retrieval. Whereas retrieval only seeks to locate similarities with provided examples, indexing takes a further step by exploiting *a priori* information to formulate classes and their respective models for classification work in a specific domain. Following is a review of the use of motion in both system types, with the latter system type being of greater relevance to our work on indexing film directing semantic concepts at the shot level.

5.2.1 Content Based Visual Query (CBVQ) for Retrieval

For the more low level works, Idris [84] used a spatiotemporal index at the shot level comprising of the spatial content in the representative frame of a shot and its camera motion. Aghbari [85] characterized every shot by the motion histograms of several representative frames. Oh [86] proposed a more elaborate scheme based on camera and object motion as well as the number of detected moving objects.

Amongst works that feature explicit and full motion segmentation, [87][88] used the long-term motion trajectories of objects for retrieval. Dagtas et al. [89] developed the most extensive work in matching motion trajectories with a video search engine called PICTURESQUE, which uses invariance spatial-temporal features for motion-based querying. Hsu [90] did similar work but modeled the trajectories using polynomials instead. Nam [91] carried out motion segmentation of video sequences

using 3-D wavelet decomposition to construct motion signatures of moving objects for storage as potential query terms. Courtney [92] tracked individual objects through the segmented data, to generate a symbolic representation of the video in the form of a directed graph. This graph describes the objects and their movement, and annotated for events of interest like appearance/disappearance and motion/rest of objects etc. However, object detection works only under the stationary case.

Chang et al proposed VideoQ [93], an object-oriented video search system capable of allowing users to specify motion trajectories and temporal duration of objects drawn in an "animated sketch" to formulate the query, which are matched to motion segmentation results for retrieval. In [94], Mezaris proposed a similar scheme and extended the previous work by utilizing intermediate-level descriptors for object attributes and inter-spatiotemporal object relationships explicitly for retrieval. Fu et al. [95] described a hierarchical approach for object-based motion description of video, which comprises of a hierarchy of low-level motions and interactions (coexist, relative directions etc.). Although Fu's framework is theoretically more detailed than Mezaris's, yet Mezaris's system is fully unsupervised, while Fu's object detection is fragile and requires manual supervision.

Fablet [80] used Gibbs models expressed in terms of co-occurrences to describe shots by certain statistical and global measures of their dynamic content, which are retrieved using the Kullback-Leibler similarity measure. Shots can be differentiated by dissimilar temporal behavior, though not sufficiently so by the spatial behavior. Furthermore it is difficult to extend the work in the total absence of the object concept.

More recently, Benini [128] used some common motion features on interesting applications like movie summarization while Chen [129] computed similiarty measures only for motion within "regions-of-interests" for retrieval purposes.

TRECVID 2005, a video retrieval evaluation track sponsored by NIST, set up the task of classifying shot-level camera motion classification into pan, tilt and zoom. The best performers (Tsinghua [99], Fudan [100] and Marburg [101]) used the most reliable motion vectors of macro-blocks provided in the MPEG stream to compute the frame-level motion type. These results are then filtered with certain rules (typically involving duration and intensity) to determine the motion type of the shot by employing parametric and statistical models to estimate the relevant motion parameters.

5.2.2 CBVQ for Indexing

In contrast to the preceding works where little or no attempt has been made to establish the semantic significance of the features used, this following group of works has explicitly mapped motion features to high level semantics. Unlike retrieval systems, existing indexing systems have hitherto chosen work in highly structured domains with more predictable and easy to model constraints (e.g. sports). This ensures the feasibility of defining a "vocabulary" of actions, or computable motion features and patterns that can map to semantics.

Lie et al [96] used simple camera motion based on affine models with a simple neural network inference engine to distinguish between 1) non-hitting, 2) in-field and 3) out-field basketball clips. Takagi [76] relied on the same camera motion models, but instead investigated the statistical properties of the camera motion type transition

parameters (pan, tilt, zoom, shake), which holds promise in distinguishing between the different genres of sports footages (baseball, soccer, tennis, sumo wrestling etc.).

Using robust motion estimation techniques, Ju et al [97] analyzed and annotated video sequences of technical talks by detecting four distinct and useful finger gestures, and segmenting out the less useful parts. Lazarescu [75] derived discrete intermediate descriptors such as camera angle, speed, number of stages in camera motion and net pan/tilt to recognize different parts of offensive plays in American football. However the assumption of the availability of hardware-supplied camera motion parameters is overly constrictive.

Ma [98] compressed the motion information of a video shot into a circular polar representation storing the magnitude and angular information of the optical flow. This "circle" is divided into four quadrants and several statistical measures such as kurtosis, skew and gyration are fed into an SVM to classify object shots (with objects), camera shots (no objects) and finally non-semantic shots (no meaningful movement). Object shots are further classified into different sports. Besides using a very arbitrary classification, there is no explicit relation between features and some classes. Rea [74] illustrated the use of the spatio-temporal behavior of an object in the footage as an embodiment of a semantic event in the snooker domain tracking the position of the white ball using a HMM with motion features to detect various snooker play semantics like shot-to-nothing, break building, conservative play and snooker escape.

Haering et al [83] proposed a three-level semantic indexing framework to detect events where the first level extracts low level cues like moving blobs while the mid-level employs a neural network to determine the object class of the blobs and generate shot descriptors. The domain-specific inference process at the third level then

uses these descriptors to detect certain user-defined events. But it is unrealistic for generic low-level features to capture domain specific information and identify object classes, especially in the general Hollywood film domain.

5.3 Semantic Taxonomy for Film Directing

Cinematography considers film as a narrative formal system where a group of interacting and interdependent film elements (sound, setting, directing etc.) are put together to deliver the narrative in a smooth and coherent manner [1]. As a vital film element, directing manifests its influence on the narrative at the shot level. With the skilful manipulation of camera motion and distance, each shot can convey additional meaning or enhanced viewer impact through accomplishing certain common yet important directing tasks. Indeed, in the light of the richness of the underlying directing semantics that can potentially be recovered, the indexing value of the directing handiwork is undeniable.

However in comparison to the well behaved directing format for sports domain, shot semantics from the film domain, at least at a sufficiently high and interesting level, are far more complex. Hence indexing directing semantics requires exploiting constraints inherent in directing grammar to construct a well-thought-out and coherent directing semantics taxonomy. Naturally, this organization should be suitably grounded upon the film directing elements.

5.3.1 Film Directing Elements

Film directing grammar encompasses two major directing elements, whose organizational relationships with shot semantics are broadly explicated and highlighted.

Camera Motion/FOA Behavior: As the chief means of subtly narrating a story through the director's perspective, camera motion in tandem with FOA (Focus of Attention) behavior, convey far more meaning than most consciously realize. Camera motion recovery is thus an indispensable step. Traditionally, the major camera motion types are pan (rotation about the vertical axis), tilt (rotation about the horizontal axis) and finally zoom (change in focal length). However for semantic level video indexing purposes, it is not so much the exact amount of zoom, or whether the camera is tilting or panning that is important, but the qualitative camera operation and its qualitative amount of movement.

In the simplest case, the *stationary* shot can be used to signify calm or even a pause pregnant with meaning. A major non-stationary shot category is the *tracking* shot, which are shots that exhibit the unique motion behavior of keeping the FOA - in other words the subject of interest - in view. Another major shot category is the *establishment* shot, whose purpose is to introduce locations instead of focusing on FOAs. Yet another major motion type with semantic significance, by virtue of its strong psychological effects on humans, is the *zoom-in*, and to a lesser extent, the *zoom-out* motion. Finally, it is recognized that not all shots are characterized by patterned camera motion and FOA behavior. Hence we create the *chaotic* shot category to describe shots that do not exhibit discernible motion patterns.

Camera Distance: Camera framing refers to the manner the FOA(s) are presented in the frame, which embodies the aspects of distance, composition, angle (low/high angled shot), level (degree of canting) and height [1]. Of these five, we concentrate on one of the most influential aspects of framing: camera distance. Cinematography admits of three coarse distance graduations (Figure 5.2): close-up, medium and long [127]. Because emotional involvement and degree of attention are approximately inversely correlated with camera distance, it can be directly used as a semantic index for the amount of attention and emotional proximity.

(a) **(b)** **(c)**

Figure 5.2 Example shots at different camera distances. Typical images of (a) close-up (b) medium and (c) long shot.

For the purpose of indexing, we have further consolidated the distance categories into two: close-ups and medium shots in one category and long shots in the other category, for the following two reasons. Firstly, on a relative basis, many more shots straddle the fine line demarcating close-ups and medium shots compared to the clean line of separation between medium and long shots: a consequence of the popular use of head-and-shoulders shots. Secondly, the camera distance is directly related to the size of the field of view, and serves as a good index of whether the director intends the shot to offer a broad overview or specific focus [104], whose distinction coincides quite neatly with the demarcation between short/medium and long shots.

5.3.2 Proposed Semantic Taxonomy

Proposing a semantic taxonomy is not an easy task, given the various criteria it should ideally fulfill. Firstly, the semantics within the taxonomy should be qualitative and meaningful to the user. For instance, commonly used low level motion classes like "pan-right" or even "tilt" should be avoided if possible, since they do not correspond strongly to higher level semantics in directing grammar and may be less relevant to the user. Secondly, the semantics selected should remain within the scope of the work as defined by its motion and film directing focus.

Since directing semantics are expressed by permutations of the directing elements, it is intuitive to utilize these directing elements as a basis to organize the semantic taxonomy. To accomplish this organization, we generate a table from all possible permutations of the directing elements (i.e. camera motion/FOA behavior and camera distance). Then we select the most meaningful and frequently employed directing semantics from directing film grammar and assign these semantics to their corresponding permutations within the table (Table 5.1).

Many advantages accrue to such an organization. The very fact that exhaustive permutations of directing elements are used as the basis of organizing the taxonomy ensures that the resultant semantic classes are relevant to the work scope and are sufficiently comprehensive to cover the whole spectrum of shots. Impossible combinations are easily detected and removed. For instance, the definition of zoom camera behavior precludes any combination with fixed camera distance. Similarly, the lack of any consistent FOA in establishment camera motion behavior renders camera distance combinations meaningless too. At the same time, feasible new semantic classes that contribute to a richer taxonomy may emerge, as in the case of Tracking,

which can logically be split into the meaningful Focus Tracking and Contextual Tracking classes along the camera distance element.

TABLE 5.1
Directing Semantics Organization by Film Directing Elements

Semantics	**Camera Motion / FOA Behavior**	**Camera Distance**
1) Stationary (Static)	**Stationary:** Little camera and FOA motion	Not Applicable: Very weak motion renders camera distance computation invalid.
2) Contextual Tracking	**Tracking:** following particular FOA	Long
3) Focus Tracking		Close Up / Medium
4) TTC (Zoom In)	**TTC:** Camera behavior dominated by camera distance	Decreasing
5) Zoom Out	**Zoom Out:** Camera behavior dominated by camera distance	Increasing
6) Intermittent / Panning Establishment	**Establishment:** Revealing surroundings and spatial relationships without particular FOA	Not Applicable: absence of FOA
7) Chaotic: Dominating sense of un-patterned motion with little coherent semantics	**Un-Patterned Motion:** Presence of large magnitude non-patterned motion from both FOA and camera	Not Applicable: Incoherent FOA and camera motion

Finally, under some circumstances, different semantics with the same permutation of directing elements can even be merged together. For instance, shots that do not belong to any of the first six semantic classes generally do not share any commonality in directing semantics, nor have any distinct directing semantics, save for the fact that they are characterized by significant un-patterned camera/FOA motions. In such a case, a new semantic class with the label "Chaotic", which accurately describes the shots motion-wise, is used to house this set of shots.

We finally settle on the seven semantic classes: 1) static, 2) contextual tracking (C-Track), 3) focus tracking (F-Track) 4) TTC, 5) zoom out, 6) establishment and 7) chaotic. Although directing semantics may necessarily be "fuzzier" and less mutually exclusive as opposed to clear-cut categories typical of the sports domain (goal shots, replay etc), clear indexing value can be fruitfully distilled from them, as the fuller descriptions of the semantic classes and their significance in the following paragraphs illustrate.

Establishment shot: In film grammar, scenes or story units should ideally start with an establishment shot (Figure 5.4 - 2nd row)[14]. Used to introduce or remind the viewer of a new environment or spatial relationships inside it, establishment shots are realized using smooth panning motion or stationary camera, with the former using panning to survey the new location. Detection of the establishment shot aids in story structure recovery such as analyzing the story units of a movie and scene segmentation. Sometimes, an establishment shot can be inserted in a long scene in order to relocate the FOA in the setting with a long shot after a number of medium and close-up shots.

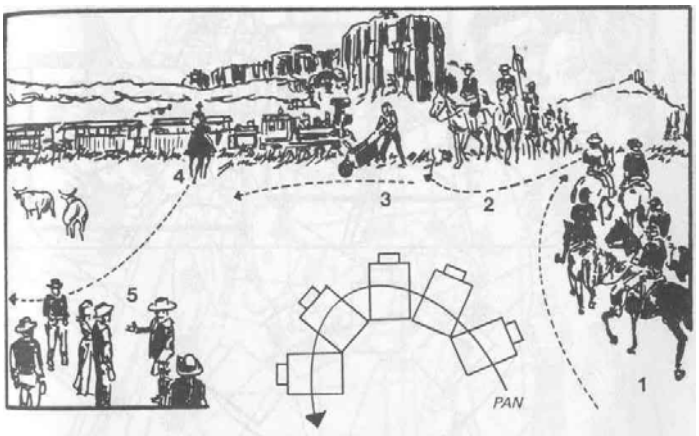


Figure 5.3 Intermittent Panning. The camera begins by following a group of soldiers moving in double line (1), then an onlooking group turning to the left (2), then a gardener (3) pushing a wheelbarrow, then a man on horseback (4) and finally the camera focuses on a group of persons talking to each other (5).

Another closely related common technique in cinematography is the intermittent pan [5], where the camera focuses the attention on multiple – usually two – FOAs by panning from one FOA to another with smooth camera movements (Figure 5.3). The chief purpose behind the intermittent pan - similar to the establishment shot - is to use the panning motion to highlight the spatial relationship between various FOAs within the shot, without focusing exclusively on any of them. Therefore it is actually a form of establishment shot, even though superficially it resembles the tracking shot.

Stationary (Static) shot: As the workhorse shot for many occasions, the vast majority of dialogue shots (over-the-shoulder, two-person shots etc.) and practically all close-up shots fall under this category [22]. The minimal motion content of such shots reflects a large portion of human interaction, which is typically sedate in terms of motion, and serves as a good index of the lull portions of a movie in contrast to its climaxes.

Tracking shot: The tracking shot is defined by a moving camera whose primary intention is to specially identify a FOA, by using the camera to either follow or rotate around the subject closely. This draws the viewer into a closer, more intense relationship with the subject [127] by creating the illusion that the viewer is directly present within the scene itself. Consequently, tracking shots are a valuable index for the strong presence of FOAs and first-person point-of-views. There exist two major variants of tracking shots. The first variant, which we call the **Focus Tracking** shot type, concentrates the viewer's attention on a subject by employing either close or medium camera shot distance. The other type, the **Contextual Tracking** shot (Figure 5.4 - 1st row), uses the long shot to show off the surroundings while accomplishing the dual purpose of tracking the subject. These two types of shots usually intertwine in longer tracking sequences as the directing requirements alternate.

Time-To-Collision (TTC) shot: A term originating from computer vision, TTC refers to the time remaining for an object to collide with the camera if the relative motion between the camera and the object remains the same. Our definition of *Time-To-Collision* shot (Figure 5.4 – 3rd row) is expanded to include shots where 1) subject or camera is moving towards the other and 2) zoom-in, where it creates an apparent impression of collision. The significance of the TTC shot lies in its ability to create and amplify emotional empathy connected with the perception of character expressions or impending collisions. This is amply demonstrated in action genre shots that use the TTC effect to create the visceral tension of impending impact (e.g. colliding car), augmenting the indexing value of TTC shots. Lastly, TTC shots can be employed to clarify details and to identify objects of importance.

Zoom-out shot: Zoom out shots, also known as detachment shots, emotionally detach or relax the interest of the viewer from the subject. This effect is usually achieved through zooming out or dolly out shots, as the camera gradually moves away from the subject and creates emotional distance. Since this shot widens the field of view, it is also employed to reveal more information about the surroundings [104].

Chaotic shot: The chaotic shot (Figure 5.4 – 4th row) refers to shots characterized by large degree of FOA movement, usually in conjunction with un-patterned camera motion. This unique motion behavior covers almost all other combinations of motion behavior not covered by other semantic classes. In this shot type, it is not unusual for the fast moving FOA to dominate viewer attention. Such shots usually cluster around the movie climax peaks and tend to be more prevalent in action genre movies.

Figure 5.4 Examples of semantic classes. Contextual tracking shot (1st row), establishment shot (2nd row), TTC shot (3rd row) and chaotic shot (4th row).

5.3.3 Shot Labeling

According to the TRECVID 2005 committee [126], in attempting to provide ground truth for shot-level camera motion to be classified into pan, tilt and zoom and no-motion, 2600 of the 5000 shots in the video corpus were rejected as being too ambiguous for annotation purposes, a situation largely attributed to 1) strong perceptual dependence on individuals and 2) presence of multiple motion types with varying strengths.

To minimize labeling subjectivity and rejection of shots from our video corpus, the choice of semantic classes is of utmost importance. Although there is still some room for judgment ambiguity in our semantics taxonomy, the more qualitative nature of our classes reduce the need to make fine quantitative judgments. For instance, under our taxonomy, it is not necessary to distinguish between pan and tilt. Neither is it required to gauge if tilt is significantly more than pan etc. Instead the primary labeling difficulty lies with the fact that a significant minority of shots have two equally significant directing semantics in consecutive arrangement within the shot (e.g. a shot with a long stationary period followed by an equally long tracking period). However identifying the two semantic classes and where they join is relatively easier, and shots felt to contain two significant directing semantics are appropriately assigned dual labels.

Additionally, we have also formulated a set of systematic guidelines below to improve objectivity in the manual assignment of ground truth for the semantic labels of some of the more ambiguous movie shots:

1) If there is a only a fragment of background framing the "tracked" foreground (the greenery outside in Figure 5.5a, then the background is totally discounted and the shot will be not labeled as a tracking shot but perhaps as a static shot if the foreground is relatively stationary.

2) Establishment shots that employ object tracking as the technique to introduce new scenery to the viewer are labeled as tracking shots. In the same vein, establishment shots filmed using a totally static shot, a technique used from time to time, are labeled as static shots (Figure 5.5b). This is in recognition that some forms of establishment

shot requires more than just motion features to detect and thus lie outside the scope of the work.

3) If a stationary camera portrays a long shot, then regardless of the FOA motion, the shot is labeled as a static shot. This is because the foreground area is simply too small to override the effects of the stationary camera (e.g. tiny fragment of FOA in the middle of Figure 5.5c).

(a) **(b)** **(c)**

Figure 5.5 Example shots to illustrate labeling rules.

5.4 Film Directing Descriptors

To distinguish shots between various directing semantics classes, the descriptors computed should intuitively possess some general relationships with the directing elements (i.e. camera motion/FOA behavior and distance) responsible for generating the semantics. This is not to claim a rigid one-to-one correspondence between directing elements and semantics. However occasional violations of film grammar do not invalidate the broad relationships between film directing elements and directing semantics. On the contrary, it argues for designing computable motion-based descriptors shown by directing grammar to possess conceptual linkages with the semantic classes. We now detail the various mid-level modules used to compute directing descriptors from the outputs of the motion segmentation module (Chapter 4), illustrated in a flowchart in Figure 5.6.

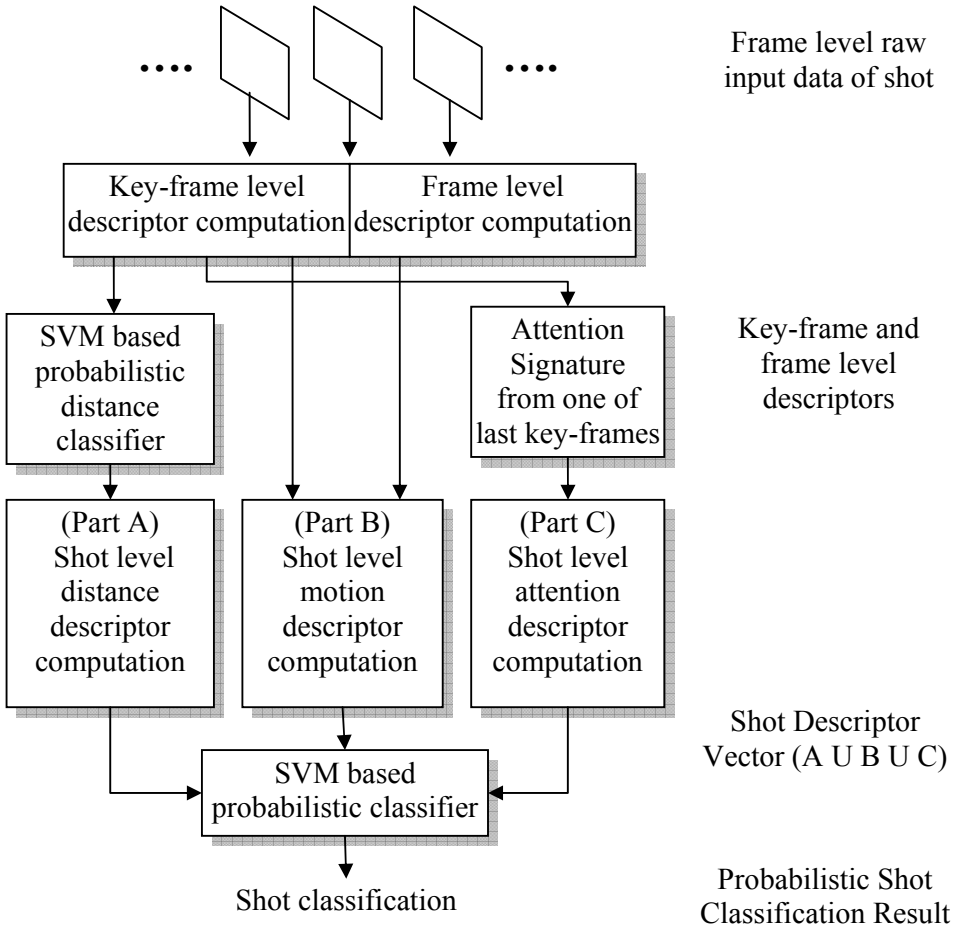


Figure 5.6 Flowchart of the shot semantics classification process.

5.4.1 Key-Frame and Frame Level Descriptors

From the motion segmentation process described in Chapter 4, every shot is represented by a number of key-frames. Starting with the first frame as a key-frame, subsequent key-frames are selected when the frame differencing threshold between the current frame under consideration and the immediate previous key-frame exceeds a fixed threshold. Thus for every key-frame, relative to the next key-frame, we are able to obtain four separate types of motion-based information. These are the 1) dense optical flow, the 2) binary background/foreground image segmentation map, the 3)

attention signature image map and finally the 4) background motion, which is succinctly represented in the affine parametric form (Equation 4.4).

Since the raw motion-related information is extracted at the key-frame level, it is necessary to normalize both the background affine parameters and optical flow distribution, *w.r.t.* the number of frames between each pair of key-frames. This is accomplished by extrapolating and smoothing the affine parameters and optical flow across the entire shot for every frame. After normalization, the background affine parameters and optical flow become shared descriptors at both the key-frame and frame level.

Now, let every key-frame be represented by a vector of descriptors, which we call the Key-Frame Descriptor Vector (KFDV). These descriptors, which are computed from the above-mentioned key-frame motion information, are as follows:

Background Speed: The background speed, or the camera motion speed, for every key-frame k in a shot is computed by $Mag_{BG,k}=\sqrt{a_1^2+a_4^2}$.

Compressed Foreground Magnitude Histogram (CFMH): Our motion segmentation algorithm computes the optical flow for the foreground, as denoted by the binary background/foreground image segmentation map, in every key-frame. This is in contrast to other works that usually exploit the motion vectors provided by the MPEG format, which are meant to minimize inter-frame differences and do not necessarily conform to the true optical flow our algorithm is designed to recover. The foreground pixels of every key-frame are represented with a polar representation histogram comprising of 8 equi-angular bins of 45 degrees and 16 motion magnitude

bins ([0.25, 1.0, 2.0, 3.0, 4.0, 5.0, 6.25, 7.75, 9.25, 11.0, 13.0, 15.0, 17.5, 20.5, 23.5, 30]), for a total of 128 bins. To obtain the final Compressed Foreground Magnitude Histogram, the 128 bin histogram is collapsed along the angle dimension and compressed along the magnitude dimension into only 5 bins of the following configuration: [1^{st}-3^{rd} bins, 4^{th}-6^{th} bins, 7^{th}-9^{th} bins, 10^{th}-12^{th} bins and 13^{th}-16^{th} bins]. Higher foreground magnitudes have a loose correspondence with closer camera distances.

Motion Vector Entropy: This measure computes the entropy of the 128 bin version of the foreground polar representation histogram. To a certain extent, the entropy admits an indirect measure of the likelihood of disparate objects in the frame, which in turn affects our inference of the number of objects, and hence the shot distance.

Foreground Area Percentage: This descriptor measures the total percentage of pixels designated as foreground *w.r.t* the frame area, and functions as the chief measure of camera distance. It is certainly true that the percentage of foreground is not strictly inversely proportionate to camera shot distance. However this descriptor functions adequately as a differentiator between the coarse camera distance categories of close up/medium shots and long shots.

5.4.2 Shot Level Distance Based Descriptor

In order to gauge the camera distance of every shot, the key-frame level descriptors of the shot, or KFDV, are fed into a probabilistic SVM classifier. This SVM classifier is trained as a 2-class classifier, using the close up/medium and long

shots as the two different classes. The percentages of key-frames in a shot that are classified into the two classes are used as the shot level distance based descriptor. This descriptor is the chief means to distinguish between contextual tracking and focus tracking shots, which differ only in their camera distance.

5.4.3 Shot Level Motion Based Descriptors

Normalized Shot Duration: This descriptor is derived by dividing the absolute duration of a shot (seconds) by $2t_{sd}$, where t_{sd}=4.3s is the average shot duration. Although semantics do not have any strict rules in film grammar concerning shot duration, chaotic shots tend to be of much shorter duration compared to most other semantic classes. On the other hand, both establishment and tracking shots tend to have a relatively long minimal duration, because both semantics require time to allow the viewers to be familiarized with the location or the tracked FOA respectively.

Stationarity Percentage: This descriptor measures the percentage of frames in the shot where Mag_{BG}<2. As opposed to other semantic classes, stationary shots tend to overwhelmingly cluster around high values of this measure.

Zoom-In and Zoom-Out Percentages: The Time-To-Collision (TTC) value for every frame is computed as $(1/(a_3+a_6))$, and TTC values between [0,400] and [-400,0] are deemed significant indicators that the camera is experiencing the zoom-in and zoom-out phenomenon respectively. Thus the percentages of frames within a shot deemed to undergo the zoom-in and zoom-out phenomenon are the Zoom-In and Zoom-Out

Percentages respectively. This is the main measure to differentiate between zooming and non-zooming shots.

Smoothness Percentage: This descriptor calculates the longest consecutive period of the camera motion being either the pan (a_1) or tilt (a_4) without a change in direction, where the camera is deemed to be in motion if $Mag_{BG}>2$. This descriptor serves as a measure of the smoothness of motion typical of establishment and tracking shots.

Shot Compressed Foreground Magnitude Histogram: This descriptor is the average CFMH of every key-frame within a shot. Very intense histograms are a good sign of the chaotic semantic class, while the converse is true for static shots.

5.4.4 Shot Level Attention Based Descriptors

Distinguishing between the establishment and tracking semantic classes, which share the same smooth background motion, requires the detection of the presence of an FOA. The attention image map (Figure 5.7), whose intensity at every pixel records the number of times it is classified as part of an FOA, serves to indicate presence of FOAs using motion-based information.

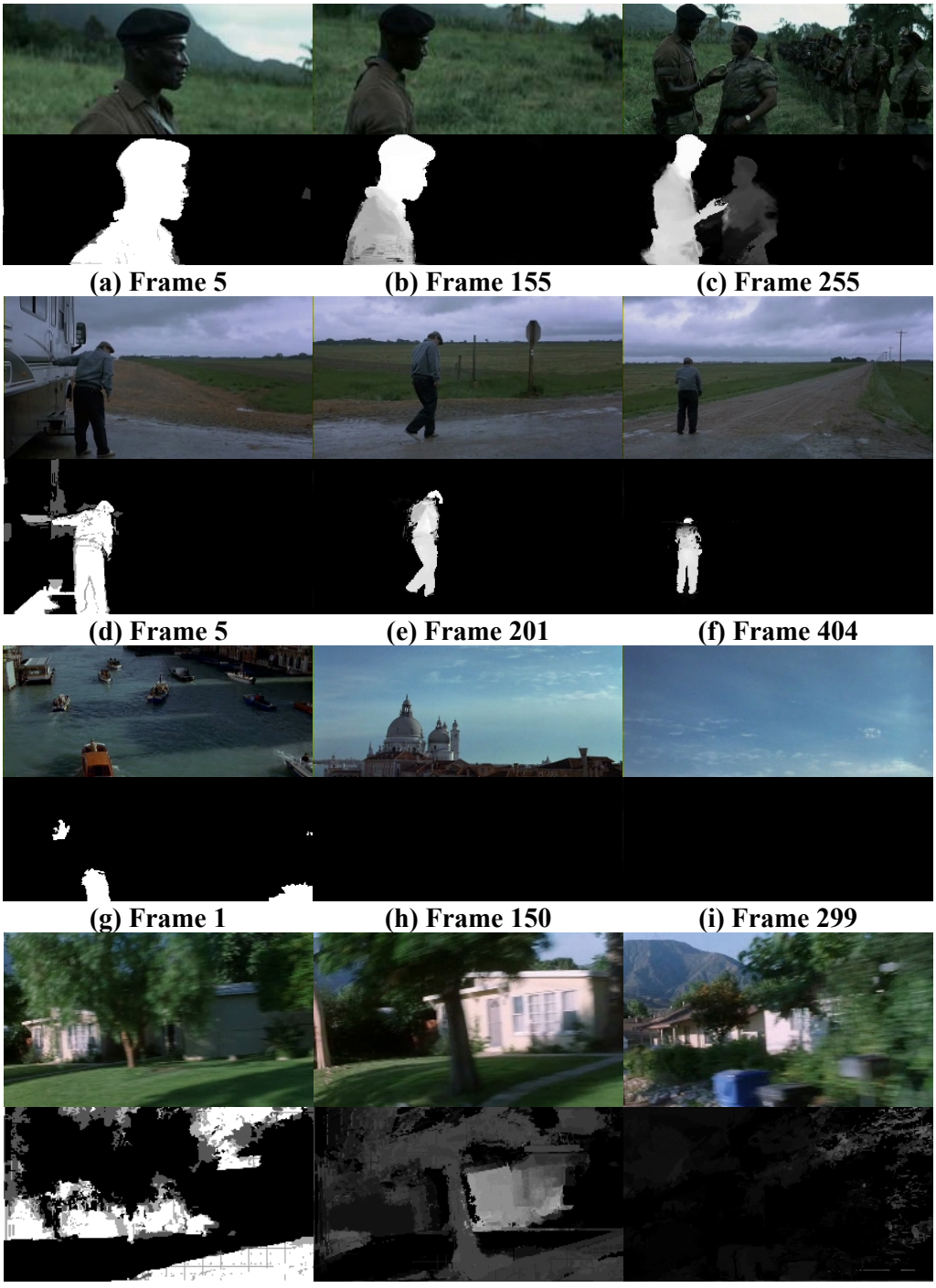
First, we normalize the values of the attention image map against the total number of key-frames in a shot. This normalized map is used to construct an equally-spaced 10 bin attention histogram $\mathbf{AH}=\{ah_1,ah_2,\ldots,ah_{10}\}$ where each ah_n denotes the proportion of pixels in the attention image map with normalized attention values falling within the bin (e.g. ah_5 will have a bin range of [0.5,0.6]). In theory, the attention histogram at the last key-frame of the sequence should be able to give a good

indication of FOA presence. However there is a noticeable tendency for some tracking shots to allow the tracked FOA to either leave the frame or be occluded in the last moments. This would destroy the attention trail that had hitherto been maintained and give spurious classification results if only the last frame were used.

To counter this problem, we compute an **AH** intensity measure, $\mathbf{AH}_{im}$, for each key-frame in the last-third portion of the shot as $\mathbf{AH}_{im} = \sum_{i}^{10} \text{median}(ah_i) * ah_i$. For instance, bin ah_5 bin has the range [0.5,0.6] and hence will have a median value of 0.55. The **AH** with the highest $\mathbf{AH}_{im}$ is finally used as the shot level attention based descriptor. This ensures the FOA is at least consistently tracked until at least the last third of a shot to fulfill the tracking criterion, and can increase robustness against occlusions that occur in the last portion of the shot.

5.4.5 Shot Descriptor Vector

The shot level attention, distance and motion descriptors of each shot are concatenated into a 21 dimension Shot Descriptor Vector (SDV) to describe the shot.



(a) Frame 5 **(b) Frame 155** **(c) Frame 255**

(d) Frame 5 **(e) Frame 201** **(f) Frame 404**

(g) Frame 1 **(h) Frame 150** **(i) Frame 299**

(j) Frame 3 **(k) Frame 37** **(l) Frame 65**

Figure 5.7 Attention signatures from four sequences (row-wise). The tracking shots are (a-c) and (d-f), whereas the establishment shots are (g-i) and (j-l). Notice that the contrast in the intensity of the attention signatures between the last frames of the tracking and establishment sequences.

5.5 Experimental Results

Our original video corpus (Table 5.2) comprises of 5226 shots lasting 366 minutes and spans across seven movies of diverse genres: two full romantic comedy movies (There's something about Mary, Bedazzled), one melodrama (City of Angels) and finally selected fast action scenes from four action movies (Lord Of the Ring I, Star Wars, James Bond and Starship Troopers). From this video corpus, we have taken out 172 extremely low intensity shots whose average frame intensities are below 30, on the basis that these shots pose problems even for manual foreground and background segmentation. Other than this one condition, the video corpus has been chosen to maximize variety. For labeling purposes two persons are employed to

TABLE 5.2
Video Corpus Description by Shot and Frames

Movie	**Shots (final/original)**	**Frames**	**Duration**
There's something about Mary	1004/1039	155938	104 mins
Bedazzled	980/1012	119641	80 mins
City of Angels	1053/1141	149545	100 mins
Lord Of the Ring I	653/659	43211	29 mins
Star Wars	495/502	27992	19 mins
James Bond	565/568	37950	21 mins
Starship Troopers	304/305	23171	13 mins
Total	5054/5226	557448	366 mins

TABLE 5.3
Composition of Directing Semantic Classes in Video Corpus (%)

	Static	ZoomOut	TTC	Estab	C-Track	F-Track	Chaotic
Number	1931	34	231	146	412	879	1421
(%)	38.21	0.67	4.57	2.89	8.15	17.39	28.12

independently label all shots according to guidelines in Chapter 5.3.3. Finally, the few labeling discrepancies between both label sets are harmonized after discussion between the labelers. The number of shots with dual labels consists of 15.3% of the entire video corpus.

To carry out shot semantics classification, the probabilistic SVM classifier (Chapter 3.4) is used to classify the Shot Descriptor Vector (SDV) (Chapter 5.4.5) that represents each shot, and outputs a 7x1 vector where each entry denotes the probability of the shot belonging to a particular shot semantic. We used a multi-class C-SVM with radial basis function kernel, with the penalty parameters C=2 and margin γ=5. To reduce the deleterious effects of the great imbalance of samples between certain classes as shown in Table 5.3 (Zoom Out and Estab classes have relatively fewer samples), we conduct training and classification using the following method.

For each training-classification iteration, we select shots for training and testing in a manner similar to bagging [132]. With this method, we randomly select from each class the smaller number between 300 and 85% of all shots of that class for training, and reserve the remaining shots for testing, allowing a much more balanced training set. During the testing phase, the test label of each test shot is the class receiving the highest probability, and the test results are noted down. This training-classification process is iterated 100 times. Finally, the test label results of every shot are tallied over all iterations and each shot is finally assigned to the test label receiving the most votes according to the tallied results. The classification results are shown in the following tables.

TABLE 5.4
Confusion Matrix for Directing Semantic Classes (%)

	Static	ZoomOut	TTC	Estab	C-Track	F-Track	Chaotic
Static	91.94	0.20	0.78	0.73	0.67	0.34	5.34
ZoomOut	2.37	83.59	1.85	0.74	5.93	1.48	2.74
TTC	0.39	0.33	87.55	0.59	0.52	2.16	8.56
Estab	2.58	0.00	1.97	82.12	4.32	3.79	3.03
C-Track	0.98	0.29	0.68	6.89	87.94	5.54	0.82
F-Track	0.62	6.36	2.43	1.12	4.96	86.45	4.07
Chaotic	6.12	0.29	1.74	0.91	1.33	4.25	85.36

TABLE 5.5
Recall and Precision for Directing Semantic Classes (%)

	Static	ZoomOut	TTC	Estab	C-Track	F-Track	Chaotic
Recall	91.94	83.59	87.55	82.12	87.94	86.45	85.36
Precision	94.57	68.29	74.81	65.97	80.44	88.17	87.90

Analyzing the confusion matrix in Table 5.4, certain classification error rates stand out and deserve explanation. It is observed that the Static and Chaotic classes tend to be confused one with another. Because the main difference between these two classes is mainly the magnitude of movement, which can be easily “misjudged” due to mild viewer subjectivity, there will inevitably be a small number of “mislabeled” borderline shots. Another source of relatively high error is that of Establishment shots being mistaken as Contextual Tracking (C-Track) shots. Both types of shots are characterized by long durations of panning motion, and sometimes if the small FOA moves too fast in a Contextual Tracking shot, it is possible for the attention trail to vanish and consequently take on the appearance of an Establishment shot.

Similarly, due to the similarities in the tracking motion, both Contextual Tracking and Focus Tracking classes have the tendency to be confused one with another. From the last column, it is noticed that Chaotic class seems most prone to confusion with other classes. This is likely because Chaotic class is the most unconstrained and unstructured class both in terms of motion characteristics and camera shot distance, thus occupying a disproportionately large area in the descriptor space and increasing the likelihood of encroaching upon other classes.

From the lst row of Table 5.5, the recall rates for all classes seem satisfactory. However due to the disproportionately small sample sizes in the video corpus for the semantic classes TTC, Zoom-out and Establishment, their precision rates are extremely susceptible to false positives from other much larger classes, which though few in number, are sufficient to significantly reduce precision rates of smaller classes.

To evaluate the effectiveness of the proposed occlusion handling mechanism for the MRF based motion segmentation algorithm, the mechanism is "switched off" by adopting the hypothesis that the dominant motion is the background all the time. The new results of such a change are tabulated in Table 5.6 and Table 5.7 and compared with its counterpart results of Table 5.4 and Table 5.5. It can be observed that although there are little differences for most results, there is a rather significant drop in classification rates when occlusion handling is "switched off" for the C-Track, F-Track and Chaotic classes.

Occlusion handling is specifically formulated to identify foreground from background. Therefore it is expected to turn in better classification rates in comparison to algorithms that assume the dominant motion is always the background, especially for classes with a higher proportion of close-up shots, which tend to feature dominant

TABLE 5.6
Confusion Matrix for Directing Semantic Classes with no Occlusion Handling (%)

	Static	ZoomOut	TTC	Estab	C-Track	F-Track	Chaotic
Static	91.24	0.20	0.78	0.73	0.67	0.34	6.04
ZoomOut	2.37	83.59	1.85	0.74	5.93	1.48	2.74
TTC	0.39	0.33	87.55	0.59	0.52	2.16	8.56
Estab	2.58	0.00	1.97	82.19	6.32	3.79	3.03
C-Track	0.98	0.29	0.68	6.89	85.34	5.54	0.82
F Track	0.62	0.36	2.43	1.12	4.96	86.15	4.57
Chaotic	6.12	0.29	1.74	0.91	2.63	7.25	81.35

TABLE 5.7
Recall and Precision for Directing Semantic Classes
with no Occlusion Handling (%)

	Static	ZoomOut	TTC	Estab	C-Track	F-Track	Chaotic
Recall	91.24	83.59	87.55	82.19	85.34	86.15	81.35
Precision	94.53	68.29	74.81	64.52	77.97	83.92	86.20

foreground. As a matter of fact, close-ups are very heavily concentrated in the Static and Chaotic classes; even F-Track class is comprised mostly of medium shots. In the event of confusion between background and foreground, Static shots by virtue of their stationarity are not likely to be mislabeled as other classes, as seen from the recall rates for the Chaotic class (Table 5.7). However Chaotic shots are much more liable to be labeled as Tracking shots, especially F-Track class (seen from its precision rate in Table 5.7), due to the relatively high magnitude motion characteristics they share. It can be concluded that the occlusion handling mechanism does seem to improve semantic classification accuracy under certain circumstances.

5.6 Conclusion

Given the wealth of underlying directing semantics residing within each shot, the indexing value of such directing handiwork is undeniable. However the choice of suitable directing semantics needs to satisfy the constraints of directing grammar. We have thus proposed to organize the semantics taxonomy based on two of the vital film directing elements: camera motion/FOA behavior and camera distance, in order to construct a coherent film directing semantics taxonomy. This framework leads us to a set of well-formed directing semantic classes and a set of effective and directing-elements-related motion descriptors. Our experiments have shown that the motion-based characteristics of the directing elements within a shot are sufficient to index its directing semantics well, despite the fact that the classes themselves correspond to relatively high level and complex semantics.

For future works, it will be useful to investigate how to approximate camera distances reliably enough to tell apart close-ups from medium shots. For instance, the usage of face detectors can conceivably give useful clues to the camera distance. Extending the idea further, contextual information such as inter-shot relationships can be incorporated into the framework, paving the way for a richer semantics taxonomy. Promising avenues for further research include the use of other modalities such as image, audio and even affective information to boost the variety of semantics available for extraction.

CHAPTER VI

CONCLUSION

In this thesis, we have focused on the semantic indexing of Hollywood movie domain, a domain chosen for both the challenges and rewards presented to machine understanding and processing. We have explored this intriguing objective from the hitherto little explored film directing and affective perspectives – in other words its motion and emotion – and proposed the frameworks and algorithms to demonstrate their feasibility on real movie video corpus.

As any avid fan or student of the cinema will attest, emotion and motion are critically intertwined with the cinema, and provide cinema with much of its meaning. Therefore an important task must be to investigate how to accomplish semantic indexing of cinema from the emotional and motion perspectives. In approaching this task, we have decided to tackle the problem with a unified approach, and ground our approach on an authoritative and objective basis: film grammar. Here we briefly recap our contributions to the objective set forth for this thesis.

Contributions for the Affective Perspective:

A complementary approach has been proposed to study and develop techniques for understanding the affective content of general Hollywood movies. We laid down a set of relevant and theoretically sound emotional categories and employed a number of low level features from cinematographic and psychological considerations to estimate these emotions. We discussed some of the important issues attendant to automated

affective understanding of film. We demonstrated the viability of the emotion categories and audiovisual features by carrying out experiments on large numbers of movies. In particular, we introduced an effective probabilistic audio inference scheme and showed the importance of audio information. Finally, we demonstrated some interesting applications with the resultant affective capabilities.

Much work remains to be done in this largely unexplored field. Firstly, the wrong classification of a small proportion of scenes shows up the inherent limitation of low-level cues (especially visual) in bridging the affective gap. Therefore in the immediate future, more complex intermediate-level cues can be implemented to further improve present results. Secondly, the existence of multiple emotions in scenes requires a more refined treatment. Lastly, it is worth investigating the possibility of finer sub-partitioning of the present affective categories, as well as further scene affective vector level analysis.

Contributions for the Film Directing Perspective:

To uncover the wealth of directing semantics behind each movie shot, we have formulated a unique approach based on a pivotal observation from film grammar: namely the manipulation of the viewer's visual attention is what that ultimately defines the directing semantics of a given shot. To capture the salient information behind this attention manipulation process, we proposed an elegant and novel edge-based MRF motion segmentation technique capable of identifying the Focus-of-Attention (FOA) areas more accurately, by utilizing edge pixels to model occlusion explicitly. The elegant integrated occlusion reasoning dispenses with the need for ad-hoc occlusion detection, allowing the foreground and background to be correctly identified at the

global level without making somewhat unrealistic assumptions on the background, as do other related works in the Hollywood domain.

The motion segmentation is robust in the dynamic film environment, where key parameters adapt automatically to the shot characteristics for optimal segmentation. Furthermore, it does not make unwarranted nor restrictive assumptions on the size, number and speed of independently moving objects. Experiments show the algorithm performs satisfactorily on real life Hollywood shots, despite some difficult shot types.

To exploit the output of the motion segmentation algorithm, we have proposed a coherent film directing semantics taxonomy based on vital film directing elements (i.e. camera motion/FOA behavior and camera distance). This framework leads us to a set of well-formed directing semantic classes and a set of effective and directing-elements-related motion descriptors. Our experiments have shown that the motion-based characteristics of the directing elements within a shot are sufficient to index its directing semantics well, despite the fact that the classes themselves correspond to relatively high level and complex semantics.

One of the immediate improvements to work on is to study how to approximate camera distances reliably enough to tell apart close-ups from medium shots. Extending the idea further, contextual information such as inter-shot relationships can be incorporated into the framework, paving the way for a richer semantics taxonomy. Promising avenues for further research include the use of other modalities such as image, audio and even affective information to boost the variety of semantics available for extraction.

In the Future:

Much remains to be done in the difficult domain of film multimedia content management, processing and understanding. At this stage, most technologies are still straddling somewhere between content retrieval and indexing. Looking into the mid-term future, capabilities such as the highly precise voice-to-text language translation and eventually text-to-semantics technologies will seem likely to emerge as the cutting edge in film multimedia content understanding. On the visual front, another technology ripe for exploration would be the exciting yet daunting object recognition. Because of its potency, and probably the corresponding need for some groundbreaking advances in AI, its deployment will probably be some way off.

However back to the near future, we foresee personalized reviewing and indexing as a vital component of the entire plethora of multimedia indexing technologies in the future, and the work in this thesis will certainly be suited to play a substantial part in this scenario. Existing and new online video communities will mature. Internet Protocol Tele-Vision will start to blossom and set off a wave of unprecedented demand for video content management systems specifically tailored to search and analyze motion pictures in a customizable manner for indexing, highlighting, summarization, data-mining, automated-editing, recommendation for consumption. With this consumption driven by the general consumer, commercial vendors and niche markets, the possibilities of exploration in this field are breathtaking.

REFERENCE

[1] D. Bordwell and K. Thompson, Film Art: An Introduction, The McGraw-Hill Companies, 7th Edition, 2004.

[2] B.T. Truong, S. Venkatesh and C. Dorai, "Automatic Scene Extraction in Motion Pictures," IEEE Trans. Circuits and Systems for Video Technology, vol. 13, no. 1, pp. 5-15, 2002.

[3] E. Kijak, G. Gravier, P. Gros, L. Oisel and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues", IEEE Intl. Conf. Multimedia Expo, vol. 3, pp. 309-312, 2003.

[4] S. Moncrieff, S. Venkatesh and C. Dorai, "Horror film genre typing and and scene labeling via audio analysis," IEEE Intl. Conf. Multimedia Expo, vol. 2, pp. 193-196, 2003.

[5] N. Haering, R.J. Qian and M.I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," IEEE Trans. Circuits and Systems for Video Technology, vol. 10, no. 6, pp. 857-868, 2000.

[6] Y. Rui, A. Gupta and A. Acero, "Automatically extracting highlights for TV baseball program," ACM Multimedia, pp. 105-115, 2000.

[7] A. Salway and M. Graham, "Extracting information about emotions in films," ACM Multimedia, pp. 299-302, 2003.

[8] A. Mittal and L.F. Cheong, "Framework for synthesizing semantic-level indexes", Multimedia Tools and Applications, vol. 20, no. 2, pp. 135-158, 2003.

[9] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 143-154, 2005.

[10] H.-B. Kang, "Affective content detection using HMMs," ACM Multimedia, pp. 259-262, 2003.

[11] Z. Rasheed, Y. Sheikh and M. Shah, "On the use of computable features for film classification," IEEE Trans. Circuits and Systems for Video Technology, vol. 15, no. 1, 2005.

[12] Y. Zhai, Z. Rasheed and M. Shah, "A framework for semantic classification of scenes using finite state machines," Conf. Image and Video Retrieval, pp. 279-288, 2004.

[13] J. Monaco, How to read a film: movies, media, multimedia, Oxford University Press, 3rd edition, 2000.

[14] D. Arijon, Grammar of the film language, Silman-James Press, 1976.

[15] R.R. Cornelius, The science of emotion. Research and tradition in the psychology of emotion, Prentice-Hall, 1996.

[16] P. Ekman et al., "Universals and cultural differences in the judgments of facial expressions of emotion," Journal of Personality and Social Psychology, vol. 54, no. 4, pp. 712-717.

[17] C. E. Osgood, G. J. Suci and P. H. Tannenbaum, The measurement of meaning, University of Illinois Press, 1957.

[18] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," Journal of Research in Personality, vol. 11, pp. 273-294, 1977.

[19] P. Valdez and A. Mehrabian, "Effects of color on emotions," Journal of Experimental Psychology: General, vol. 123, no.4, pp. 394-409, 1994.

[20] J. W. Hill, Baroque music: Music in Western Europe, 1580-1750, W. W. Norton and Company, 2005.
[21] R. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, "Attention to television: Alpha power and its relationship to image motion and emotional content," Media Psychology, vol. 5, pp. 283-301, 2003.
[22] H. Zettl, Sight Sound Motion: Applied Media Aesthetics, Wadsworth Publishing Company, 3rd edition, 1998.
[23] B. Adams, C. Dorai and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo," IEEE Trans. Multimedia, vol. 4, no. 4, pp. 472-481, 2002.
[24] L. Lu, H. Jiang and H.J. Zhang, "A robust audio classification and segmentation method," ACM Multimedia, pp. 103-122, 2001.
[25] D. Liu, L. Lu and H.-J. Zhang, "Automatic mood detection from acoustic music data," ISMIR, pp. 81-87, 2004.
[26] T.L. New, S.W. Foo and L.C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, pp. 603-623, 2003.
[27] F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech," Intl. Conf. Spoken Language Processing, pp. 1970-1973, 1996.
[28] W. Chai and B. Vercoe, "Structural analysis of musical signals for indexing and thumbnailing," Intl. Conf. on Digital Libraries, pp. 27-34, 2003.
[29] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, 1998.
[30] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Microsoft Research, http://research.microsoft.com/~jplatt, 1999.
[31] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," The Annals of Statistics, vol. 26, no. 2, pp. 451-471, 1998.
[32] R.W. Picard, E. Vyzas and J. Healey, "Towards machine emotional intelligence: Analysis of affective physiological state," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175-1191, 2001.
[33] H. Schlosberg, "Three dimensions of emotion," Psychological Review, vol. 61, pp. 81-88, 1954.
[34] S. Fischer, R. Lienhart and W. Effelsberg, "Automatic recognition of film genres," ACM Multimedia, pp. 295-304, 1995.
[35] P. Ekman, "Facial Expression and Emotion," American Psychologist, vol. 48, no. 4, pp. 384-392, 1993.
[36] P. Ekman, Handbook of Cognition and Emotion: Chapter 3 (Basic Emotions), John Wiley and Sons, 1999.
[37] A. Ortony and T. Turner, "What's basic about basic emotions," Psychological Review, vol. 97, no. 3, pp. 315-331, 1990.
[38] C. Darwin, The expression of emotions in man and animals, University of Chicago Press, 1872/1965.
[39] A. Magda, Emotion and Personality, Columbia University Press, 1960.
[40] R. C. Solomon, "Back to basics: On the very idea of basic emotions," Journal for the Theory of Social Behaviour, vol. 32, no. 2, pp. 315-331, 2002.
[41] I. Fonagy and K. Magdics, "A new method of investigating the perception of prosodic features," Language and Speech, vol. 21, pp. 34-49, 1978.

[42] G. Tzanetakis and P. Cook, “Music genre classification of audio signals,” IEEE Trans. Speech Audio Processing, vol. 10, no. 5, pp. 293-302, 2002.
[43] P. N. Juslin and J.A. Sloboda, Music and emotion: theory and research, Oxford University Press, 2001.
[44] K.J. Kurtz, “Category-based similarity”. Proc. 18th Annual Conf of the Cognitive Science Society, p. 790. Hillsdale, NJ: Erlbaum. 1996.
[45] The Internet Movie Database, http://www.imdb.com/.
[46] P. Ekman, “Are there basic emotions,” Psychology Review, vol. 99, no. 3, pp. 550-553, 1992.
[47] J.-M. Odobez and P. Bouthemy, “Robust multiresolution estimation of parametric motion models”, Journal of Visual Communication and Image Representation, vol. 6, no. 4, pp. 348-365, 1995.
[48] J.-M. Odobez and P. Bouthemy, “ Direct incremental model-based image motion segmentation for video analysis”, Signal Processing, vol. 66, no. 3, pp. 143-156, 1998.
[49] J. Y. A. Wang and E. H. Adelson, “Representing moving images with layers,” IEEE Trans. Image Processing, vol. 3, pp. 625-637, 1994.
[50] Jianbo Shi and Jitendra Malik, “Normalized cuts and image segmentation”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, 2000.
[51] Jianbo Shi and Jitendra Malik, “Motion segmentation and tracking using normalized cuts”, IEEE Intl. Conf. of Computer Vision, pp. 1154-1160, 1998.
[52] H. Xu, A. A. Younis,and M. R. Kabuka, “Automatic moving object extraction for content-based applications,” IEEE Trans. Circuits and Systems for Video Technology, vol. 14, no. 6, 2004.
[53] T. Papadimitriou, K. I. Diamantaras, M. G. Strintzis and M. Roumeliotis, “Video scene segmentation using spatial contour and 3-D robust motion estimation,” IEEE Trans. Circuits Systems Video Technology, vol. 14, no. 4, 2004.
[54] V. Mezaris, I. Kompatsiaris, M. G. Strintzis, “Video object segmentation using bayes-based temporal tracking and trajectory-based region merging,” IEEE Trans. Circuits and Systems for Video Technology, vol. 14, no. 6, 2004.
[55] D. S. Tweed and A. D. Calway, “Integrated segmentation and depth ordering of motion layers in image sequences,” Image and Vision Computing, vol. 20, pp. 709-723, 2002.
[56] E. Sifakis, I. Grinias, G. Tziritas, “Video segmentation using fast marching and region growing algorithms”, EURASIP Journal on Applied Signal Processing, vol. 4, pp. 379-388, 2002.
[57] A.-R. Mansouri and J. Konrad, “Multiple motion segmentation with level sets”, IEEE Trans. Image Processing, vol. 12, no. 2, pp. 201-220, 2003.
[58] D. Cremers and S. Soatto, “Variational space-time motion segmentation”, IEEE Intl. Conf. Computer Vision, vol. 2, pp. 886-892, 2003.
[59] M. M. Chang, A. M. Tekalp, and M. I. Sezan, “Simultaneous motion estimation and segmentation,” IEEE Trans. Image Processing, vol. 6, pp. 1326-1333, 1997.
[60] P. Smith, T. Drummond and R. Cipolla, “Layered motion segmentation and depth ordering by tracking edges,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 4, pp. 479-494, 2004.

[61] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, pp. 326-332, 2001.

[62] H. S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 814-830, 1996.

[63] Y. Weiss and E. H. Adelson, "A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models", Proc. Conf. Computer Vision and Pattern Recognition, pp. 321-326, 1996.

[64] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach to layer extraction from image sequences," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, pp. 297-303, 2001.

[65] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: a region labeling approach," IEEE Trans. Circuit Systems Video Technologies., vol. 12, pp. 597-612, 2002.

[66] N. Vasconcelos and A. Lippman, "Empirical Bayesian motion segmentation," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, pp. 217-221, 2001.

[67] P. De Smet and D De Vleeschauwer, "Performance and scalability of a highly optimized rainfalling watershed algorithm," in Proc. Int. Conf. Imaging Science, Systems and Technology, vol. CISST 98, pp. 266-273, 1998.

[68] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, pp. 583-598, 1991.

[69] B. T. Truong, S. Venkatesh and C. Dorai, "Discovering semantics from visualizations of film takes," IEEE Proc. Intl. Conf. Multimedia Modeling, 2004.

[70] F. Nack and A. Parkes, "The application of video semantics and theme representation in automated video editing," Multimedia Tools and Applications, vol. 4, pp. 57-83, 1997.

[71] R Hammoud and R. Mohr, "Interactive tools for constructing and browsing structures for movie films," ACM Multimedia, pp. 497-498, 2000.

[72] N. Dimitrova and F. Golshani, "Motion recovery for video content classification," ACM Trans. Information Systems, vol. 13, no. 4, pp. 408-439, 1995.

[73] C. G.M. Snoek and Marcel Worring, "Multimodal Video Indexing: A review of the state-of-the-art," Multimedia Tools and Applications, vol. 25, no. 1, pp. 5-35, 2005.

[74] N. Rea, R. Dahyot and A. Kokaram, "Modeling high level structure in sports with motion driven HMMs," IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, pp. 621-624, 2004.

[75] M. Lazarescu and S. Venkatesh, "Using camera motion to identify types of American football plays," IEEE Proc. Int'l Conf. Multimedia and Expo, pp. 181-184, 2003.

[76] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate and H. Tominaga, "Sports video categorizing method using camera motion parameters," IEEE Proc. Int'l Conf. Multimedia and Expo, pp. 461-464, 2003.

[77] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian and C.-S. Xu, “A mid-level representation framework for semantic sports video analysis,” ACM Multimedia, pp. 33-44, 2003.

[78] S. Dagtas, W. Al-Khatib, A. Ghafoor and R. L. Kashyap, “Models for Motion-Based Video Indexing and Retrieval,” IEEE Trans. Image Processing, vol. 9, no.1, pp. 88-101, 2000.

[79] X. Sun, B. S. Manjunath, and A. Divakaran, “Representation of motion activity in hierarchical levels for video indexing and filtering,” IEEE Proc. Int’l Conf. Image Processing, pp. 149-152, 2002.

[80] R. Fablet, P. Bouthemy and P. Perez, “Nonparametric Motion Characterization Using Causal Probabilistic Models for Video Indexing and Retrieval,” IEEE Trans. Image Processing, vol. 11, no. 4, pp. 393-407, 2002.

[81] M. R. Naphade and T. S. Huang. “A probabilistic framework for semantic video indexing, filtering and retrieval,” IEEE Trans. Multimedia, pp. 141-151, 2001.

[82] C.-T. Hsu and C.-W. Lee, “Statistical motion characterization for video content classification,” IEEE Proc. Int’l Conf. Multimedia and Expo, pp. 1599-1602, 2004.

[83] N. Haering, R. J. Qian and M. I. Sezan, “A semantic event-detection and its application to detecting hunts in wildlife video,” IEEE Trans. Circuits and Systems for Video Technology, vol. 10, no. 6, pp. 857-868, 2000.

[84] F. M. Idris and S. Panchanathan, “Spatio-temporal indexing of vector quantized video sequences,” IEEE Trans. Circuits and Systems for Video Technology, vol. 7, no. 5, pp. 728-740, 1997.

[85] Z. Aghbari, K. Kaneko and A. Makinouchi, “A motion-location based indexing method for retrieving MPEG videos,” 9th Int'l. Workshop on Database and Expert Systems Applications, pp. 012-107, 1998.

[86] J. H. Oh, M. Thenneru and N. Jiang, “Hierarchical video indexing based on changes of camera and object motions,” ACM Applied Computing, pp. 917-921, 2003.

[87] Y. Jin and F. Mokhtarian, “Efficient Video Retrieval by Motion Trajectory,” British Machine Vision Conference, pp. 667-676, 2004.

[88] W. You, K. W. Lee, J.-G. Kim, J. Kim and O.-S. Kwon, “Content-based video retrieval by indexing object’s motion trajectory,” IEEE Int’l. Conf. on Consumer Electronics, pp. 352-353, 2001.

[89] S. Dagtas, W. Al-Khatib, A. Ghafoor and R. L. Kashyap, “Models for Motion-Based Video Indexing and Retrieval,” IEEE Trans. Image Processing, vol. 9, no. 1, pp. 88-101, 2000.

[90] C.-T. Hsu and S.-J. Teng, "Motion trajectory based video indexing and retrieval," IEEE Proc. Int’l Conf. Image Processing, pp. 605-608, 2002.

[91] J. Nam and A. H. Tewfik, “Progressive resolution motion indexing of video object,” IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, 1998.

[92] J. D. Courtney, “Automatic video indexing via object motion analysis,” Pattern Recognition, vol. 30, no. 4, pp. 607-625, 1997.

[93] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram and S. Zhong, “VideoQ: an automated content based video search system using visual cues,” ACM Multimedia, pp. 313-324, 1997.

[94] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris and M. G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing," IEEE Trans. Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 606-621, 2004.
[95] Y. Fu, A. Ekin, A. M. Tekalp and R. Mehrota, "Temporal segmentation of video objects for hierarchical object-based motion description," IEEE Trans. Image Processing, vol. 11, no. 2, 2002.
[96] W.-N. Lie, T.-H. Lin and S.-H. Hsia, "Motion-based event detection and semantic classification for baseball sports video," IEEE Proc. Int'l Conf. Multimedia and Expo, pp. 1567-1569, 2004.
[97] S. X. Ju, M. J. Black, S. Minnemant and D. Kimbert, "Analysis of gesture and action in technical talks for video indexing," IEEE Proc. Computer Vision and Pattern Recognition, pp. 595-601, 1997.
[98] Y.-F. Ma and H.-J. Zhang, "Motion pattern based video classification using support vector machines," IEEE Int'l Symposium Circuits and Systems, pp. 69-72, 2002.
[99] J. Yuan, H. Wang, L. Xiao, D. Wang, D. Ding, Y. Zuo, Z. Tong, X. Liu, S. Xu, W. Zheng, X. Li, Z. Si, J. Li, F. Lin and B. Zhang, "Tsinghua University at TRECVID 2005," TRECVID 2005.
[100] X. Xue, L. Hong, L. Wu , Y. Guo, Y. Xu, C. Mi, J. Zhang, S. Liu, D. Yao, B. Li, S. Zhang, H. Yu, W. Zhang and B. Wang, "Fudan University at TRECVID 2005," TRECVID 2005.
[101] R. Ewerth, C. Beringer, T. Kopp, M. Niebergall, T. Stadelmann and B. Freisleben, "University of Marburg at TRECVID 2005: Shot Boundary Detection and Camera Motion Estimation Results," TRECVID 2005.
[102] L.-Y. Duan, J. Wang, Y. Zheng, C. Xu, Q. Tian, "Shot-level camera motion estimation based on a parametric model," TRECVID 2005.
[103] "The MPEG-7 visual part of the XM 4.0, ISO/IEC MPEG99/W3068," HI, 1999.
[104] S. D. Katz, Film directing shot by shot: visualizing from concept to screen, Michael Wiese Productions, 1991.
[105] A .L. Gasskill and D. A. Englander, How to shoot a movie and video story: The technique of pictorial continuity, Morgan and Morgan Inc. Publishers, 1985.
[106] S. Pfeiffer, R. Lienhart, G. Kühne and W. Effelsberg, "The MoCA project: Movie content analysis research at the university of Mannheim," Informatik, pp. 329-338, 1998.
[107] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, A. C. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval, IEEE Trans. Multimedia, vol. 7, no. 4, 2005.
[108] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, 2000.
[109] Internet Movie Data Base, http://www.imdb.com/, 2005.
[110] N. Dimitrova, L. Agnihotri, and G. Wei, "Video classification based on HMM using text and faces," European Signal Processing Conference, 2000.
[111] V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replay, text, and camera motion features," in SPIE Conference on Storage and Retrieval for Media Databases, vol. 3972, pp. 332–343, 2000.

[112] B.T. Truong, S. Venkatesh, and C. Dorai, "Automatic genre identification for content-based video categorization," IEEE Intl. Conf. Pattern Recognition, vol. 4, pp. 230-233, 2000.
[113] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: naming and detecting faces in news videos," IEEE Multimedia, vol. 6, no. 1, pp. 22–35, 1999.
[114] S. Eickeler and S. Muller, "Content-based video indexing of TV broadcast news using hidden markov models," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2997–3000, 1999.
[115] L. Bergen and F. Meyer; "A novel approach to depth ordering in monocular image sequences," IEEE Intl. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 536-541, 2000.
[116] C. Stiller, "Object-based estimation of dense motion fields," IEEE Trans. Image Processing, vol. 6, no. 2, pp. 234-250, 1997.
[117] S. Geman and D.Geman, "Stochastic relaxation, gibbs distribution and the Bayesian restoration of images, " IEEE Trans. Pattern Analysis and Machine Intelligence vol. 6, no. 6, pp. 721-741, 1984.
[118] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," Intl. Journal of Computer Vision, vol. 4, pp. 185-210, 1990.
[119] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," Journal Royal Statistics Society B, vol. 36, no. 2, pp. 192-236, 1974.
[120] L. Y. Duan, J. S. Jin, Q. Tian, C.-S. Xu, "Nonparametric motion characterization for robust classification of camera motion patterns," IEEE Trans. Multimedia, vol. 8, no. 2, pp. 323-340, 2006.
[121] Y.-F. Ma, X.-S. Hua, L. Lu and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. Multimedia, vol. 7, no. 5, pp 907- 919, 2005.
[122] P. Bouthemy, M. Gelgon and F. Ganansia, "A unified approach to shot change detection and camera motiion characterization," IEEE Trans. Circuits and Systems for Video Technology, vol. 9, no. 7, pp. 1030-1044, pp. 34-47.
[123] W. B. Thompson, "Combining motion and contrast for segmentation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 2, no. 6, pp. 543-549, 1980.
[124] N. Peyrard and P. Bouthemy, "Motion-based selection of relevant video segments for video summarization," Multimedia Tools and Applications, vol. 26, pp. 259-276, 2005.
[125] Y.-H. Ho, C.-W. Lin, J.-F. Chen and H.-Y. M. Liao, "Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics," IEEE Trans. Circuits and Systems for Video Technology, vol. 16, no. 5, pp. 642-648, 2006.
[126] P. Over, T. Ianeva, W. Kraaijz and A. F. Smeaton, "TRECVID 2005 - An Overview", TRECVID 2005, 2006.
[127] J. S. Douglass and G. P. Harnden, The Art of Technique: An Aesthetic Approach to Film and Video Production, Allyn & Bacon, 1st Edition, 1995.
[128] S. Benini, L.-Q. Xu and R. Leonardi, "Using Lateral Ranking for Motion-Based Video Shot Retrieval and Dynamic Content Characterization," Fourth International Workshop on Content-Based Multimedia Indexing (CBMI), 2005.

[129] J.-F. Chen, H.-Y. M. Liao and C.W. Lin, "Fast Video Retrieval via the Statistics of Motion Within the Regions-of-Interest," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 437-440, 2005.
[130] Kolmogorov-Smirnov Goodness-of-Fit Test, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm, 2007.
[131] Z. Li Stan, Markov Random Field Modeling in Image Analysis, Springer-Verlag, 1st edition, 1995.
[132] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123-140, 1996.
[133] K. Wyatt and C.Schroeder, Harmony and Theory: A Comprehensive Source for All Musicians, Music Press Institute, 1st edition, 1998.